Journal of Experimental Psychology: Applied
2009, Vol. 15, No. 1, 1–11

© 2009 American Psychological Association
1076-898X/09/$12.00 DOI: 10.1037/a0014721

Memorial Consequences of Answering SAT II Questions

Elizabeth J. Marsh
Duke University

Pooja K. Agarwal and Henry L. Roediger, III
Washington University in St. Louis

Many thousands of students take standardized tests every year. In the current research, we asked whether answering standardized test questions affects students' later test performance. Prior research has shown both positive and negative effects of multiple-choice testing on later tests, with negative effects arising from students selecting incorrect alternatives on multiple-choice tests and then believing they were correct (Roediger & Marsh, 2005). In the current experiments, undergraduates and high school students answered multiple-choice questions retired from SAT II tests (that are no longer in the testing pool) on biology, chemistry, U.S. history, and world history, and later answered cued-recall questions about these subjects. In 3 experiments, we observed positive testing effects: More final cued-recall questions were answered correctly if the items had appeared on the initial multiple-choice test. We also sometimes observed negative testing effects: intrusions of multiple-choice distractors as answers on the final cued-recall test. Students who scored well on the initial test benefited from taking the test, but lower achieving students showed either less benefit (undergraduates) or costs from the testing (high school students).

Keywords: testing effect, multiple-choice tests, negative suggestion effect

A rite of passage for high school students is taking standardized tests such as the SAT and the ACT. These tests are used for assessment purposes to measure general aptitude (the SAT reasoning test) or achievement in specific subject matters (the ACT, the SAT subject tests). Because the tests often play a determining role in important admission and placement decisions, many students complete practice examinations, read books about the tests, or enroll in expensive preparatory classes to learn how to perform well on them. The usefulness and predictive validity of these and other similar tests are important issues both in American education and abroad (e.g., Kobrin, Patterson, Shaw, Mattern, & Barbuti, 2008; Kuncel, Hezlett, & Ones, 2001).

Our interest is not in the predictive validity of such standardized tests but rather in the consequences of taking such tests on students' performance on later tests. In that sense, our focus is similar to that espoused in the formative assessment literature, where tests paired with feedback are treated as learning opportunities for both the student and the teacher (e.g., Black & Wiliam, 1998). However, our experiments examined learning from tests never paired with feedback. We hypothesized that simply answering standardized questions (such as those from the SAT) may change the way students respond on later tests. This hypothesis is based on the large literature developed in both laboratory and educational settings that reveals that tests not only measure what is learned, but can also change students' performance in both positive and negative directions (see Roediger & Karpicke, 2006a, 2006b).

The *testing effect* refers to the finding that taking a test on recently studied material generally improves students' performance on a later test (e.g., Bjork, 1975; Glover, 1989; Hogan & Kintsch, 1971; McDaniel & Masson, 1985). Testing not only measures retention but also changes the accessibility of information on later tests, often (but not always) in a positive manner. Although the educational implications of testing effect studies are sometimes noted (Gates, 1917; Glover, 1989; McDaniel & Fisher, 1991; Spitzer, 1939), these studies have not permeated educational practice. The research summarized here refers to classroom testing or students' self-testing (e.g., via flash cards), but we can ask the same question with regard to standardized questions: Will answering SAT II questions affect later test performance?

As noted above, testing has also been shown to have negative effects. The reason is, in a way, simple: Because students learn from tests, if the tests contain erroneous information, students may learn errors. Whereas educators would never consider embedding wrong information in their lectures or assigned readings, they routinely use testing methods that expose students to misinformation. Both true-false and multiple-choice tests provide erroneous information, and the research reviewed below shows that both types of tests may have detrimental effects on students' later performance, what Remmers and Remmers (1926) termed the *negative suggestion effect*.

Each SAT II question (which assesses knowledge of a specific domain such as chemistry) is a multiple-choice question

Elizabeth J. Marsh, Department of Psychology and Neuroscience, Duke University; Pooja K. Agarwal and Henry L. Roediger, III, Department of Psychology, Washington University in St. Louis.

This research was supported by a collaborative activity award from the James S. McDonnell Foundation and Grant R305H060080-06 to Washington University in St. Louis from the Institute of Education Sciences, U.S. Department of Education. The opinions expressed are those of the authors and do not represent the views of the James S. McDonnell Foundation, the Institute, or the U.S. Department of Education. We are grateful to principal Sara Keene and guidance counselor Doug Rose of Columbia High School for helping us recruit the high school subjects. We also thank Kristy Duprey for her assistance with Experiment 3, and Holli Sink and Barbie Huelser for their help preparing materials and testing participants.

Correspondence concerning this article should be addressed to Elizabeth J. Marsh, Department of Psychology and Neuroscience, Duke University, 9 Flowers Drive, Durham, NC 27708-0086. E-mail: emarsh@psych.duke.edu

that includes five possible answers: Four are incorrect and only one is correct.[1] Yet, carefully reading all alternatives to evaluate them necessitates students' exposure to four alternatives containing errors (or misinformation). However, simply reading statements (even if labeled as false) can make them seem true on a later assessment, an outcome known as the *illusory truth effect* (Bacon, 1979; Hasher, Goldstein, & Toppino, 1977). This research suggests that even if students select the correct answer on a multiple-choice test, reading the wrong alternatives may make them later seem true. Supporting this idea, Toppino and colleagues showed that distractors (incorrect alternatives) from multiple-choice and true-false exams are later judged as truer than novel false facts (Toppino & Brochin, 1989; Toppino & Luipersbeck, 1993; see also Rees, 1986), although the rated truth of repeated falsehoods never reached the level of objectively true statements. We found that prior reading of a greater number of multiple-choice distractors (varied across conditions from one to five) decreased the positive testing effect and increased production of multiple-choice distractors as incorrect answers on the final test (Butler, Marsh, Goode, & Roediger, 2006; Marsh, Roediger, Bjork, & Bjork, 2007; Roediger & Marsh, 2005). In our work, negative testing effects occurred after multiple-choice distractors had been selected on the initial test; because the SAT II test employs four distractors, the negative effects of testing might be great and may even eliminate any positive effect of testing.

To address these issues, we used retired SAT II materials that are in the public domain (and are no longer in the testing pool) to see whether answering these questions would yield positive or negative effects as in prior research (Roediger & Marsh, 2005). Because the SAT II employs many distractors and because students often receive no feedback on the correctness of their responses other than a summary score, the negative suggestion effect from testing may outweigh the positive effects.

On the other hand, there are two critical facts about the SAT II that may inoculate test-takers against learning falsehoods. First, the SAT II penalizes test-takers for wrong answers; its scoring system should discourage guessing. This feature is important given that we previously have found that the multiple-choice test's negative effects were limited to distractors that were selected rather than those merely read (Fazio, Agarwal, Marsh, & Roediger, 2008; Roediger & Marsh, 2005). Second, the SAT II taps relatively high-level knowledge about domains, and usually only those students who feel they have mastered the subject matter reasonably well take the tests. By high-level knowledge, we mean the questions are not simple definitional ones, but rather involve evaluating, analyzing, and applying concepts—and as such tap higher levels in Bloom's (1956) taxonomy of educational objectives. This combination of a penalty for wrong answers, more complex questions, and well-prepared students may not yield the same negative suggestion effects as we have observed with simpler materials.

In short, it is an open question as to whether answering SAT II questions will benefit or harm students' later test performance. To determine the answer, we tested subjects with retired SAT II materials and examined the consequences of such testing on a later cued-recall test covering the same material. To preview, positive and negative testing effects were observed in the first study, and the two follow-up studies were designed to better understand the negative effects of testing. In the second study, we manipulated the instructions on the multiple-choice test to see whether instructions that encouraged subjects to select more multiple-choice distractors would increase the negative suggestion effect. The first two experiments involved Duke University undergraduates, who took at least three SAT II tests as part of the application process. On average, Duke undergraduates score highly on standardized tests; the middle 50% range of SAT scores of students accepted for the class of 2011 was 680–770 for the reading section and 690–790 for math. The third experiment involved less experienced test-takers: high school students at a suburban school in Illinois. This group was expected to select more multiple-choice distractors than would the undergraduates, allowing us a second way to vary performance on the initial multiple-choice test. Across the two populations, the question was how answering SAT II questions would affect performance on a later test.

Experiment 1

Method

Participants. Thirty-two Duke undergraduates participated in the study[2]; they received either course credit or $10.

Materials. Twenty 5-alternative multiple-choice questions were selected from each of four SAT II tests (College Board, 2002): biology, chemistry, United States history, and world history. Questions were selected only if they could be tested in cued-recall format, which was simply an open-ended question (in other words, the question prompt was presented without the five response options and subjects typed each response into a text box). For example, a question might probe the consequences of the Nazi–Soviet Nonaggression Pact of 1939. In the multiple-choice version, subjects had to discriminate among five possible answers, such as the destruction of the League of Nations (incorrect) and allowing Germany and the USSR to attack Poland without interference from each other (correct). The selected questions were similar in average difficulty (62% answered correctly on the real SAT II) to the entire set in the SAT II practice book (56.5% correct).

Each set of 20 questions per domain was randomly split into two sets matched for difficulty; across subjects, we counterbalanced which set was tested on the initial test and which was omitted. All 80 questions (40 previously tested, 40 not tested) were used on the final cued-recall test. As with the real SAT II, questions were blocked by domain and the order of domain was counterbalanced across subjects; within each subtest, questions were tested in random order for each subject.

Design. The study had a 2 (multiple-choice test: tested or not tested) × 4 (domain: biology, chemistry, U.S. history, world history) within-subjects design.

[1] The SAT II tests have been renamed the SAT subject tests, but to our knowledge the name change was not accompanied by changes in formatting or scoring.

[2] We did not collect demographics, but the sample is expected to reflect the Duke undergraduate population, which is 49% female and 51% male.

Procedure. All testing was computerized.[3] Subjects were instructed to pretend that they were taking real SAT II tests, and the SAT II scoring system was explained. Subjects learned they would earn 1 point per correct answer, *lose* 1/4 point for each error, and that skipping questions would not affect points. They were instructed to earn the best score possible on each of the mini-SAT II tests, and that they should answer strategically to reach that goal.

For each SAT II question, subjects selected one of the five possible answers or skipped the question. There were 40 questions on the initial test, 10 from each domain. Following a 5.25-min filled delay involving visuospatial puzzles, subjects answered the final 80 cued-recall questions. They were explicitly instructed that the SAT scoring system no longer applied. However, they were warned against guessing and instructed to type "I don't know" if they did not know an answer.

Results

All results were significant at the .05 level unless otherwise noted.

Performance on the initial multiple-choice test. On average, the subjects answered 55% of the SAT II test questions correctly ($SD = 15$). As shown in Table 1, performance was highest for biology questions and lowest for chemistry, with the two history tests intermediate; the main effect of domain was significant, $F(3, 93) = 5.97$, $MSE = 0.04$, $\eta_p^2 = .16$. The same pattern was obtained when the dependent measure was the SAT score (number correct – 1/4 point per incorrect answer) rather than proportion correct, $F(3, 93) = 5.23$, $MSE = 5.11$, $\eta_p^2 = .14$.

An interesting result is that distractors were equally likely to be selected across the four domains ($M = 22\%$, $SD = 11$, $F < 1$). It is these items that might give rise to a negative suggestion effect on the final test.

Final test: Correct performance. As on the initial test, the subjects again answered more biology questions correctly than they did for the other domains, $F(3, 93) = 2.70$, $MSE = 0.05$, $p = .05$, $\eta_p^2 = .08$. More important, a significant testing effect was observed in all four domains, although there was a trend for the testing effect to be slightly larger in some domains than others, $F(3, 93) = 2.46$, $MSE = 0.02$, $p < .07$, $\eta_p^2 = .07$. Overall, as shown in Table 2, a large positive testing effect was obtained: Subjects correctly answered more cued-recall questions if the questions had appeared on the prior multiple-choice test ($M = 0.48$, $SD = 0.17$) than if they had not ($M = 0.22$, $SD = 0.10$), and this testing effect was significant, $F(1, 31) = 138.96$, $MSE = 0.03$, $\eta_p^2 = .82$.

Table 1
Distribution of Responses on the Initial SAT II Questions as a Function of Academic Domain, Experiment 1

Response	Biology	Chemistry	World history	U.S. history	*M* (*SE*)
Correct	0.66	0.44	0.54	0.58	0.55 (0.027)
Distractor	0.21	0.23	0.21	0.22	0.22 (0.020)
Skipped	0.13	0.33	0.25	0.20	0.23 (0.021)

Table 2
Proportion of Final Cued-Recall Questions Answered Correctly (Top Panel) or With SAT II Multiple-Choice Distractors (Bottom Panel), Experiment 1

Response	Biology	Chemistry	World history	U.S. history	*M* (*SE*)
Correct					
Tested	0.55	0.40	0.44	0.51	0.48 (0.029)
Not tested	0.27	0.21	0.20	0.20	0.22 (0.018)
Difference	0.28	0.19	0.24	0.31	0.26 (0.022)
Distractors					
Tested	0.15	0.20	0.15	0.14	0.16 (0.016)
Not tested	0.05	0.12	0.07	0.04	0.07 (0.007)
Difference	0.10	0.08	0.08	0.10	0.09 (0.016)

Final test: Production of SAT II distractors. A negative suggestion effect was obtained with SAT II materials. As shown in the bottom portion of Table 2, production of the multiple-choice distractors as answers on the final test was higher for questions tested on the initial exam ($M = 0.16$, $SD = 0.09$) than for questions that had not been tested previously ($M = 0.07$, $SD = 0.04$), $F(1, 31) = 33.05$, $MSE = 0.02$, $\eta_p^2 = .52$. Although multiple-choice distractor answers were more frequent for some of the domains than others, $F(3, 93) = 5.36$, $MSE = 0.01$, $\eta_p^2 = .15$, the effect of testing was similar across domains ($F < 1$).

For tested items, two additional analyses examined the relationship between multiple-choice distractor selection and final cued-recall answers. First, given that a distractor was selected on the multiple-choice test, what was the probability of it persisting to the final test? Sixty-three percent of selected distractors were also produced on the final cued-recall test. Recall that no feedback was given after the multiple-choice test and subjects were told to be conservative on the final cued-recall test. Therefore, errors that persevered were likely ones the subjects believed to be correct. Second, given that a multiple-choice distractor was produced on the final cued-recall test, what was the likelihood that it had been selected on the initial multiple-choice test? Over 80% of multiple-choice distractors produced on the final test were the same as the answer that had been selected on the initial multiple-choice test. In other words, testing did not lead subjects to reproduce read but nonselected distractors on the final test.

Final test: Production of other errors and questions left unanswered. Keeping in mind that the four response categories (correct, SAT II distractor, other wrong answer, and "don't know") were not independent, we describe briefly how testing affected the production of other wrong answers and "don't know" responses. Testing reduced "don't know" responses, $F(1, 31) = 34.08$, $MSE = 0.038$, $\eta_p^2 = .52$, and also reduced the production of other wrong answers, $F(1, 31) = 116.81$, $MSE = 0.023$, $\eta_p^2 = .79$. Although testing *increased* the production of SAT distractor an-

[3] Note that computerized testing did *not* mean that the questions were tailored to the examinee's ability level (computer-adaptive testing was not used). On the initial multiple-choice test, one half of students answered the same set of 40 questions regardless of their performance (and the other half of students answered the other 40 questions).

swers on the final test, it also *decreased* the production of other wrong answers and *increased* the production of correct answers. As shown in the first pair of bars in Figure 1, although SAT distractor intrusions on the final test were 9% higher for previously tested items, the overall error rate (multiple-choice distractors plus other wrong answers) still declined from 33% to 22%. Testing significantly reduced the overall error rate on the final cued-recall test, $F(1, 31) = 29.75$, $MSE = 0.03$, $\eta_p^2 = .49$.

Discussion

Answering SAT II questions yielded a positive testing effect: Subjects answered more questions correctly on the final cued-recall test if they had previously been tested on the initial multiple-choice test. Although answering SAT II questions also increased production of the multiple-choice distractors on the final test, it did not, however, increase the overall error rate. That is, whereas SAT II testing increased multiple-choice distractor answers on the final test, this effect was accompanied by a decrease in other errors. Therefore, taking the SAT II led to a net gain in performance on the cued-recall test.

One comment is warranted on the effect of domains: Not all domains were equally easy on the multiple-choice test, and correspondingly the positive testing effect was stronger in some domains than others. In other words, average multiple-choice performance in a domain was correlated with the size of the positive testing effect ($r = .82$). Testing was beneficial to the extent that subjects were able to retrieve and select the answers. Distractor selection was constant across domains, and similarly the negative testing effect was consistent across domains.

Some standardized tests require cautious responding and penalize wrong answers, and others (e.g., the ACT) do not. Previous work documenting negative consequences of testing (e.g., Roediger & Marsh, 2005) has used forced responding, whereas the SAT II does not. Experiment 2 was conducted to examine the contribution of forced responding to the negative suggestion effect. We included a condition to directly replicate Experiment 1 and added a new condition that required subjects to answer every question to examine whether forced responding would increase the negative suggestion effect on the final test.

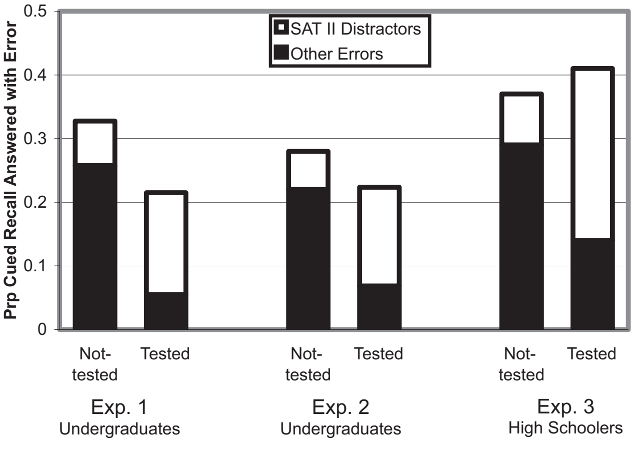

Figure 1. Production of SAT II distractors and other wrong answers as a function of testing. The subjects were undergraduates in Experiments 1 and 2 and high school students in Experiment 3.

Experiment 2

Method

Participants. Undergraduates ($N = 103$) from the same population as Experiment 1 participated in the study for either course credit or $10. Subjects were excluded from the analyses if their multiple-choice performance was more than 2 standard deviations from the mean; 7 subjects were excluded using this criterion. Thus, the analyses contain data from 96 subjects, with all factors completely counterbalanced.

Materials. All materials were the same as in Experiment 1, except for the multiple-choice test. The SAT II questions were always paired with the same five alternatives, but only subjects in the free-responding condition were given the option to skip questions (as in Experiment 1).

Design. The experiment employed a 2 (multiple-choice test: tested or not tested) × 4 (Domain: biology, chemistry, U.S. history, world history) × 2 (multiple-choice responding: free vs. forced) mixed design. Multiple-choice testing and domain were manipulated within subjects, whereas type of responding on the initial test was varied between subjects.

Procedure. Procedures were the same as in Experiment 1, except for the manipulation of instructions for the initial multiple-choice test: Subjects received either free- or forced-responding instructions. In the forced-responding instructions, subjects were told to select one answer for each question, even if they had to guess. Free-responding instructions were the same as the SAT II instructions used in Experiment 1; subjects were informed there was no penalty for skipping questions, but that they would lose 1/4 point for each error and gain 1 point for each correct answer. On the final test, as in Experiment 1, subjects were urged to be cautious and permitted to skip answering questions by writing “I don’t know” beside the query.

Results

Performance on the initial multiple-choice test. The data are shown in Table 3. A 4 (domain: biology, chemistry, U.S. history, world history) × 2 (multiple-choice responding: free vs. forced) mixed ANOVA was computed on proportions correct. As in Experiment 1, there was a main effect of domain, $F(3, 282) = 11.63$, $MSE = 0.039$, $\eta_p^2 = .11$, indicating that some question types were easier than others. However, the important finding was that response instructions neither affected correct responding nor interacted with domain (Fs < 1). Subjects in the forced-responding condition answered no more questions correctly ($M = 0.58$, $SD = 0.13$) than did subjects in the free-responding condition ($M = 0.57$, $SD = 0.14$).

However, as expected, forced-responding instructions increased errors on the initial multiple-choice test, $F(1, 94) = 70.58$, $MSE = 0.059$, $\eta_p^2 = .43$. When subjects were forced to select one of the five possible answers, they selected twice as many distractors ($M = 0.42$, $SD = 0.13$) as did subjects who were allowed to skip questions ($M = 0.21$, $SD = 0.12$). Forced responding increased distractor selection in all domains, although this effect was larger in some domains than others, $F(3, 282) = 3.51$, $MSE = 0.027$, $\eta_p^2 = .04$.

Table 3
Distribution of Responses on the Initial SAT II Questions as a Function of Academic Domain, Experiment 2

Response	Biology	Chemistry	World history	U.S. history	*M* (*SE*)
Free responding					
Correct	0.64	0.50	0.55	0.57	0.57 (0.019)
Distractor	0.20	0.23	0.22	0.20	0.21 (0.018)
Skipped	0.16	0.28	0.23	0.23	0.22 (0.021)
Forced responding					
Correct	0.67	0.49	0.56	0.61	0.58 (0.019)
Distractor	0.33	0.51	0.44	0.39	0.42 (0.018)

Of primary interest are the effects of forced responding on later test performance, as examined in the next sections.

Final test: Correct performance. As shown in Table 4, a large positive testing effect was obtained: Across the two conditions, subjects correctly answered a greater proportion of cued-recall questions if they had been tested on the prior multiple-choice test (M = 0.45, SD = 0.16) than if they had not (M = 0.23, SD = 0.10), F(1, 94) = 297.06, MSE = 0.032, η_p^2 = .76. A significant testing effect was observed in all four domains, although the testing effect was slightly larger in some domains than others, F(3, 282) = 10.08, MSE = 0.02, η_p^2 = .10. Consistent with the findings on the multiple-choice test, there was no difference in proportion correct on the final test between free- (M = 0.36, SD = 0.12) and forced-responding (M = 0.33, SD = 0.12) conditions, F(1, 94) = 1.80, MSE = 0.117, $p > .18$, η_p^2 = .02. More importantly, the positive testing effect was equally large regardless of whether free or forced responding was required when answering SAT II questions; the interaction between response options and testing was not significant, F(1, 94) = 1.53, MSE = 0.03, $p > .2$, η_p^2 = .02.

Final test: Production of SAT II distractors. The data in Table 5 show that subjects answered more final questions with SAT II distractors if the items had previously been tested (M = 0.16, SD = 0.09) than if they had not been tested (M = 0.06, SD = 0.04), F(1, 94) = 122.85, MSE = 0.014, η_p^2 = .57, thus replicating the negative suggestion effect obtained in Experiment 1. The effect held for all domains tested, and the effect of domain did not interact with testing, F(3, 282) = 1.25, MSE = 0.01, $p > .29$, η_p^2 = .01.

Most importantly, after testing, erroneous responding with SAT II distractors increased to 13.8% in the free-responding condition and 17.3% in the forced-responding condition; the baseline (not tested) production of SAT II distractors was 6% in both conditions. In other words, forced responding led to greater intrusions of tested SAT II distractors on the final test than did free responding. Although the effect is small, it was observed across all four subject domains, and the interaction between testing and responding was significant, F(1, 94) = 4.65, MSE = 0.01, η_p^2 = .05.

Table 4
Proportion of Final Cued-Recall Questions Correctly Answered (Experiment 2) as a Function of Multiple-Choice Responding Instructions, Prior Multiple-Choice Testing, and Domain

Response	Biology	Chemistry	World history	U.S. history	*M* (*SE*)
Free responding					
Tested	0.56	0.44	0.42	0.48	0.48 (0.024)
Not tested	0.28	0.28	0.20	0.20	0.24 (0.014)
Difference	0.28	0.16	0.22	0.28	0.24 (0.019)
Forced responding					
Tested	0.49	0.35	0.40	0.46	0.43 (0.023)
Not tested	0.26	0.25	0.19	0.19	0.22 (0.016)
Difference	0.23	0.10	0.21	0.27	0.21 (0.017)

Table 5
Proportion of Final Cued-Recall Questions Answered With Multiple-Choice Distractors (Experiment 2) as a Function of Multiple-Choice Responding Instructions, Prior Multiple-Choice Testing, and Domain

Response	Biology	Chemistry	World history	U.S. history	*M* (*SE*)
Free responding					
Tested	0.13	0.16	0.16	0.10	0.14 (0.011)
Not tested	0.04	0.09	0.06	0.07	0.06 (0.005)
Difference	0.09	0.07	0.10	0.03	0.08 (0.010)
Forced responding					
Tested	0.16	0.19	0.18	0.16	0.17 (0.013)
Not tested	0.04	0.09	0.07	0.04	0.06 (0.006)
Difference	0.12	0.10	0.11	0.12	0.11 (0.014)

For tested items, two additional analyses examined the relationship between multiple-choice distractor selection and final cued-recall answers. First, given that a multiple-choice distractor was selected on the initial test, what was the likelihood of it being reproduced on the final cued-recall test? In the free-responding condition, selected distractors persisted to the final cued-recall test at a rate (M = 0.61, SD = 0.27) very similar to that observed in Experiment 1 (M = 0.63, SD = 0.23), which also allowed free responding. However, distractors were more likely to persist to the final test if they were selected in the free-responding condition (M = 0.61, SD = 0.27) than the forced-responding condition (M = 0.39, SD = 0.22), t(94) = 4.58, SE = 0.05. In other words, forced responding led to a much higher rate of multiple-choice distractor selection, but these selections were less likely to persist to the final cued-recall test, presumably because many of these were guesses.

Similar to Experiment 1, 89% of the distractors that appeared on the final test had been selected on the prior SAT II test, and this did not differ between the free- (M = 0.88, SD = 0.18) and forced-responding conditions (M = 0.90, SD = 0.11), $t < 1$.

Final test: Production of other errors and questions left unanswered. Once again, keeping in mind that the four response categories (correct, SAT II distractor, other wrong answer, and "don't know") were not independent, we describe briefly how testing affected the production of "don't know" and other wrong answers. Testing reduced both "don't know," F(1, 94) = 140.20, MSE = 0.037, η_p^2 = .60, and other wrong responses on the final test, F(1, 94) = 219.18, MSE = 0.020, η_p^2 = .70. As in Experiment 1, even though testing *increased* production of SAT II distractors on the final test, the accompanying *decrease* in other wrong answers meant that testing did significantly reduce the total number of errors, F(1, 94) = 20.20, MSE = 0.03, η_p^2 = .18. This pattern of data is depicted in the second pair of bars in Figure 1.

Discussion

The data from Experiment 2 nicely parallel the findings of Experiment 1: Subjects correctly answered a greater proportion of cued-recall questions if the items had been tested on the initial mock SAT II test. This positive testing effect was again stronger in some domains than others; domains associated with better multiple-choice performance yielded larger positive testing effects ($r = .78$). Although testing did increase intrusions of multiple-choice distractors on the final cued-recall test, this negative testing effect was much smaller than the observed benefits of testing. Distractor selection did not vary across domains, and consequently there were no domain differences in the negative suggestion effect.

Of particular interest are the data from subjects in the forced-responding condition. These subjects selected many more distractors than did subjects in the free-responding condition, yielding a small but significant increase in multiple-choice distractor answers on the final test. On the one hand, forced responding did not seem costly: It did not affect the positive testing effect and only increased distractor intrusions by 3%. On the other hand, forced responding did significantly increase the negative testing effect (even if only by 3%), hinting that the cost may be larger in other testing situations where the negative suggestion effect is larger than observed here. Another possibility is that the benefits of free responding might increase if the penalty for errors were increased: Subjects lost only 1/4 point for errors here, meaning that they should guess if they could eliminate just one of the five response options. With a larger penalty, they might be less likely to select distractors, with consequences for the negative suggestion effect.

One slightly puzzling result in Experiment 2 was that forced responding did not increase the proportion of items answered correctly on the multiple-choice test, with a difference of only 1% in the expected direction (58% to 57%). Of course, we may not have had the power to detect a difference, but one might have expected that forcing subjects to guess would have increased correct responding, even if just by chance (e.g., Koriat & Goldsmith, 1996). However, this did not seem to occur; rather, when forced to respond, subjects seemingly picked only distractor items. Although this outcome is puzzling, it does demonstrate that our subjects showed exquisite metacognitive control for these difficult items (roughly 22% of those tested). They skipped the items because they believed they did not know enough to answer them correctly and, when we forced a comparable group to respond, we showed that they were right—even with forced responding, they generally failed to pick the correct answer and improve their scores. Of course, their metacognitive awareness was not perfect, because they were wrong in answering about 20% of the questions.

In both experiments described thus far, the subjects were Duke University undergraduates, who are on average high scorers on standardized tests (and thus not prototypical of the students who take these kinds of exams every year). But even in this relatively homogeneous sample, post hoc analyses suggested individual differences in the memorial consequences of testing. We reexamined positive and negative testing effects as a function of how subjects scored on the initial multiple-choice test. We rank-ordered Experiment 2 subjects by multiple-choice scores (collapsed over domain) and compared the final cued-recall performance of subjects who scored in the top 25% on the multiple-choice test (the top 24 subjects) with those who scored in the bottom 25% (the bottom 24 subjects). Table 6 shows the size of the testing effects for the top and bottom performers; note that the data from Experiment 2 appear in the middle of the table. Subjects in the top 25% showed a robust positive testing effect on the final cued-recall test, increasing from a not-tested baseline of 31% correct to 65% for items that had been tested on the prior multiple-choice test. Subjects in the bottom 25% also showed a positive testing effect, increasing from 16% to 27%. However, their net increase was only 11%, whereas the net increase from testing was 34% for top subjects. This difference resulted in a significant interaction between testing and multiple-choice performance, $F(1, 46) = 61.04$, $MSE = 0.01$, $\eta_p^2 = .57$. In contrast, the highest ability subjects were less likely to reproduce multiple-choice distractors on the final cued-recall test following multiple-choice testing. Both groups intruded similar levels of multiple-choice distractors for nontested items (0.07 for both groups). For previously tested items, however, subjects from the bottom 25% intruded more multiple-choice distractors on the cued-recall test ($M = 0.19$, $SD = 0.10$) than did subjects from the top 25% ($M = 0.14$, $SD = 0.07$). In other words, the interaction between multiple-choice performance and testing was significant

Table 6
Positive and Negative Testing Effects, for Students in Each Experiment Who Scored in the Bottom 25% and Top 25% on the Initial Multiple-Choice Test

	Correct answers		Distractor answers	
Variable	Bottom 25%	Top 25%	Bottom 25%	Top 25%
Experiment 1				
Tested	0.37	0.69	0.20	0.08
Not tested	0.12	0.32	0.08	0.08
M difference (*SE*)	0.25 (0.034)	0.37 (0.034)	0.12 (0.029)	0.00 (0.021)
Experiment 2				
Tested	0.27	0.65	0.19	0.14
Not tested	0.16	0.31	0.07	0.07
M difference (*SE*)	0.11 (0.019)	0.34 (0.023)	0.12 (0.021)	0.07 (0.013)
Experiment 3				
Tested	0.18	0.35	0.26	0.29
Not tested	0.09	0.13	0.09	0.11
M difference (*SE*)	0.09 (0.028)	0.22 (0.023)	0.17 (0.034)	0.18 (0.021)

for the negative suggestion effect as well, $F(1, 46) = 5.03$, $MSE = 0.01$, $\eta_p^2 = .10$. Critically, the two groups showed opposite patterns: The highest scoring subjects showed a larger positive testing effect and a smaller negative suggestion effect, whereas the lowest scoring subjects showed a smaller positive testing effect and a larger negative testing effect. This larger negative suggestion effect in the lowest performers means that these subjects no longer showed the overall reduction in the error rate depicted in Figure 1. That is, multiple-choice performance significantly interacted with the effects of testing on the total error rate, $F(1, 46) = 23.18$, $MSE = 0.01$, $\eta_p^2 = .34$. The highest achievers saw a drop in the overall error rate from 0.31 ($SD = 0.12$) to 0.18 ($SD = 0.09$) after testing, $t(27) = 6.47$, $SE = 0.02$, because of the large drop in other wrong answers that accompanied the small increase in distractor intrusions. In contrast, the overall error rate did not change as a function of testing for the subjects with the lowest multiple-choice scores, $t(27) = 1.12$, $SE = 0.03$, $p > .27$.

We reanalyzed the data from Experiment 1 to see whether the same patterns were obtained, keeping in mind that there were only 8 subjects per group (the 32 subjects were rank-ordered by multiple-choice performance, and the subjects in the top 25% and bottom 25% were identified). These data are in the top panel of Table 6. The positive testing effect was larger for high performers than the low performers, $F(1, 14) = 7.11$, $MSE = 0.005$, $\eta_p^2 = .34$, and the high performers were also less likely to show a negative suggestion effect on the final test, $F(1, 14) = 11.39$, $MSE = 0.003$, $\eta_p^2 = .45$. Testing reduced total error rate (multiple-choice distractors and other wrong answers) in both groups, although the drop was larger for high performers than low performers, $F(1, 14) = 3.94$, $MSE = 0.01$, $p < .07$, $\eta_p^2 = .22$. The pattern of results parallels those observed in Experiment 2. Combined, the data emphasize the need to examine the memorial consequences of testing for students who do not perform as well. Testing is unlikely to benefit low performers as much as it helps high performers.

To obtain a broader range of multiple-choice scores, in Experiment 3 we tested high school students at a public school in Illinois. The school is located in a suburban bedroom community of St. Louis, with a 99% graduation rate; approximately 60% of the graduates matriculate at 4-year universities and about 30% go on to 2-year schools. The graduating class of 2007 received an average ACT composite score of 22, which is comparable to an SAT score of 1,030 out of a possible 1,600. In contrast, as noted earlier, the middle 50% of Duke's class of 2011 scored much higher: ACT scores ranged from 29 to 34, and SAT scores from 1,370 to 1,560. Thus, we did not expect these high school students to do as well as the undergraduates on the initial multiple-choice test, and we wanted to see the memorial consequences of this poorer initial performance. We used exactly the same materials and procedures as in Experiment 1 (except for a switch to paper-and-pencil testing) to conduct planned group comparisons between the high school students and the Experiment 1 undergraduates.

Experiment 3

Method

Participants. Twenty-eight subjects (18 females; ages 16–17 years; M age = 16.3 years) from an Illinois public high school participated in a single group session, and each subject received $10 for participating. The subjects were juniors who were recruited by the school guidance counselor from a larger group of juniors participating in a practice PSAT session. Parental consent and written assent from each subject were obtained. One subject was excluded from analyses because his multiple-choice performance was more than 2 standard deviations from the mean. Thus, the analyses contain data from 27 subjects.

Materials. All materials were the same as in Experiment 1, except that paper-and-pencil format was used. Subjects recorded their multiple-choice test responses on a Scantron answer sheet, similar to the real SAT II, and final cued-recall test responses were recorded in the test booklet. As in Experiment 1, sets of questions included and omitted from the initial test were counterbalanced. Questions were blocked by domain, the order of which was also counterbalanced across subjects. Within each block and counterbalancing condition, however, questions were tested in a fixed random order because of the paper-and-pencil format.

Design. This study had the same design as Experiment 1: a 2 (multiple-choice test: tested or not tested) × 4 (domain: biology, chemistry, U.S. history, world history) within-subjects design.

Procedure. Procedures were similar to those in Experiment 1. The initial multiple-choice test was self-paced and a 27-min time limit was provided, although all subjects completed the test within this time limit. As in Experiment 1, an explanation of the SAT II scoring system was given; participants learned they would earn 1 point per correct answer, lose 1/4 point per error, and that skipped questions would not affect points. Following a 5.25-min filled delay, the final cued-recall test was self-paced. The length of the delay, the contents of the filler task, and the instructions for final cued recall were all identical to Experiment 1.

Results

Performance on the initial multiple-choice test. On average, subjects answered 34% ($SD = 0.10$) of the SAT II questions correctly. As shown in Table 7, performance was greatest for the biology domain, with similar performance in the chemistry, world history, and U.S. history domains. As in Experiments 1 and 2, the main effect of domain was significant, $F(3, 78) = 3.12$, $MSE = 0.02$, $\eta_p^2 = .11$. This pattern did not change when analyses were conducted using SAT II scores (number correct – 1/4 point per incorrect answer) instead of proportion correct. In addition, distractors were selected equally often across the domains, 56%, $F(3, 78) = 1.35$, $p > .25$.

Final test: Correct performance. Subjects correctly answered more cued-recall chemistry and biology questions than history questions, $F(3, 78) = 6.40$, $MSE = 0.02$, $\eta_p^2 = .20$. As demonstrated in Experiments 1 and 2 and shown in Table 8, a signif-

Table 7
Distribution of Responses on the Initial SAT II Questions as a Function of Academic Domain, Experiment 3

Response	Biology	Chemistry	World history	U.S. history	*M* (*SE*)
Correct	0.42	0.32	0.31	0.32	0.34 (0.020)
Distractor	0.52	0.57	0.57	0.60	0.56 (0.024)
Skipped	0.06	0.11	0.11	0.08	0.09 (0.029)

Table 8
Proportion of Final Cued-Recall Questions Answered Correctly (Top Panel) or With SAT II Multiple-Choice Distractors (Bottom Panel) as a Function of Prior Multiple-Choice Testing and Domain, Experiment 3

Response	Biology	Chemistry	World history	U.S. history	*M* (*SE*)
Correct					
Tested	0.29	0.26	0.21	0.21	0.24 (0.020)
Not tested	0.10	0.20	0.10	0.05	0.12 (0.011)
Difference	0.19	0.06	0.10	0.16	0.12 (0.017)
Distractors					
Tested	0.24	0.34	0.29	0.24	0.28 (0.018)
Not tested	0.04	0.16	0.06	0.06	0.08 (0.011)
Difference	0.20	0.18	0.23	0.18	0.20 (0.019)

icant testing effect was obtained, such that subjects correctly answered a greater proportion of questions if they were previously tested on the initial multiple-choice test ($M = 0.24$, $SD = 0.10$) than if they were not ($M = 0.12$, $SD = 0.06$), $F(1, 26) = 55.13$, $MSE = 0.02$, $\eta_p^2 = .68$. This testing effect, however, varied in size across domains, $F(3, 78) = 3.69$, $MSE = 0.01$, $\eta_p^2 = .12$. The positive testing effect was stronger for domains associated with higher multiple-choice performance ($r = .70$).

In Experiments 1 and 2, the highest performing subjects showed a larger positive testing effect than did the subjects who performed the worst on the initial multiple-choice test. This same analysis was completed for Experiment 3; the seven highest scoring subjects were compared with the seven lowest scoring subjects. The results are shown in the bottom panel of Table 6. Although both groups showed positive testing effects, this effect was stronger in the best performers than in the worst performers, $F(1, 12) = 13.97$, $MSE = 0.002$, $\eta_p^2 = .54$. The advantage of tested items over baseline was 22% for high performers but only 9% for the low performers.

Final test: Production of SAT II distractors. The negative suggestion effect found in Experiments 1 and 2 was replicated in Experiment 3: Subjects produced a greater proportion of SAT II distractors on cued-recall questions that were previously tested in comparison to questions that were not tested, $F(1, 26) = 102.88$, $MSE = 0.02$, $\eta_p^2 = .80$. These data are in the bottom portion of Table 8. This negative suggestion effect was true of all four domains, and the effect of domain did not interact with the effect of testing, $F < 1$.

We also compared multiple-choice distractor intrusions for the top and bottom performers. However, in this case, there was no difference in the size of the negative suggestion effect between groups, $F < 1$. As shown in Table 6, both groups of subjects showed robust negative testing effects.

Finally, two additional analyses were done on previously tested items to link multiple-choice and final cued-recall performance. First, we examined the likelihood of a multiple-choice distractor persisting, given that it was selected on the initial test. Over 40% of multiple-choice distractor selections were reproduced on the final test. Second, 82% of the distractors subjects produced on the final test were the same as those selected on the initial multiple-choice test.

Final test: Production of other errors and questions left unanswered. Similar to both Experiments 1 and 2, testing reduced skipped answers, $F(1, 26) = 35.22$, $MSE = 0.04$, $\eta_p^2 = .58$, as well as other incorrect answers on the final test, $F(1, 26) = 74.49$, $MSE = 0.018$, $\eta_p^2 = .74$. In other words, although testing increased the proportion of distractors on the final test, other wrong answers and "don't know" responses were reduced. To see the effect of testing on the overall error rate, we summed the multiple-choice distractor and other wrong responses and redid the ANOVA. In contrast to the two earlier experiments, whereby testing significantly reduced the overall error rate, here the effect of testing was not significant, $F(1, 26) = 2.47$, $MSE = 0.035$, $p < .13$, $\eta_p^2 = .09$. It is intriguing that the effect was in the *opposite* direction of what we observed earlier, with a trend toward a higher total error rate after testing ($M = 0.42$, $SD = 0.14$) than for nontested items ($M = 0.38$, $SD = 0.18$). Testing further interacted with academic domain, $F(3, 78) = 4.32$, $MSE = 0.02$, $\eta_p^2 = .14$. Post hoc analyses revealed two significant effects in world history and chemistry. In both cases, testing led to a greater total error rate than was observed for nontested items, $t(26) = 2.44$, $SE = 0.04$, and $t(26) = 2.73$, $SE = 0.04$. There were no significant differences between previously tested and nontested items for biology and U.S. history (*t*s < 1).

Comparison of high school and undergraduate students. As described earlier, the use of the same materials and procedures allowed the comparison of the high school students with the undergraduates from Experiment 1. We collapsed over academic domain and compared overall performance on the multiple-choice and cued-recall tests across groups.

Compared with the undergraduates, the high school students answered fewer multiple-choice questions correctly, $t(57) = 6.14$, $SE = 0.03$, selected far more distractors, $t(57) = 10.08$, $SE = 0.03$, and did not skip as many questions, $t(57) = 4.34$, $SE = 0.03$. When the undergraduates did not answer a multiple-choice question correctly, half the time they answered it and half the time they skipped it (see Table 1). In contrast, when the high school students did not answer a multiple-choice question correctly, they selected a distractor 86% of the time (see Table 7). In addition to answering more multiple-choice questions correctly, the undergraduates showed much more refined metacognitive awareness of what they did versus did not know.

To examine the memorial consequences of the differences in multiple-choice performance, we computed a 2 (multiple-choice test: tested, not-tested) × 2 (group: undergraduate, high school) ANOVA on proportion of cued-recall questions answered correctly. As already described, both groups showed a positive testing effect. The important point for present purposes is that the positive testing effect was larger in the undergraduates than in the high school students; the interaction between testing and group was significant, $F(1, 57) = 20.36$, $MSE = 0.01$, $\eta_p^2 = .26$. Testing improved performance by 26% in the undergraduates but only 12% in the high school students (see Tables 2 and 8), and this difference was significant, $t(57) = 4.51$, $SE = 0.03$.

In contrast, testing led to a larger negative suggestion effect in the high school students than in the undergraduates, $F(1, 57) = 18.11$, $MSE = 0.01$, $\eta_p^2 = .24$. As shown in Tables 2 and 8, the two groups intruded similar levels of multiple-choice distractors for items that had not previously been tested (0.07 for the undergraduates and 0.08 for the high school students). Testing increased multiple-choice distractor intrusions in both groups, but high school students increased 20% over their baseline, whereas undergraduates increased only 9%, $t(57) = 4.26$, $SE = 0.02$.

Finally, did the relationship between the multiple-choice and cued-recall tests differ across groups? Recall that 31 of the 32 undergraduates in Experiment 1 selected distractors on the multiple-choice test, and of those 63% persisted to the final test. In contrast, for high school students, only 43% of the selected multiple-choice distractors were reproduced on the final cued-recall test, and this difference was significant, $t(56) = 3.53$, $SE = 0.06$.

Discussion

High school students answered the same SAT II questions as did undergraduates in Experiment 1 under exactly the same instructions. On one level, the effects were similar across the two groups: Both high school students and undergraduates showed positive testing effects and the negative suggestion effect. For both populations, the positive testing effect was larger for domains associated with better multiple-choice performance. However, undergraduates showed a larger positive testing effect and a smaller negative suggestion effect than did the high school students. Undergraduates clearly benefited from answering SAT II questions; after testing, they answered more questions correctly, and the increase in SAT II distractor intrusions was balanced by a large drop in other wrong answers, as shown in Figure 1.

The conclusions about testing high school students are less positive. Although they did answer more questions correctly following testing, their positive testing effect was smaller than that observed in undergraduates. The high school students also intruded more SAT II distractors after testing, and in contrast to the undergraduates, this increase was not balanced by a similar decrease in other wrong answers. Figure 1 shows that, if anything, testing increased the overall error rate in high school students. Statistically, in two domains there was no effect of testing on the overall error rate, and in two domains testing actually *increased* the overall error rate. Therefore, although testing also led to a decrease in other wrong answers in high school students, the increase in SAT II distractor intrusions was so large that the overall error rate slightly increased after testing.

General Discussion

In all three experiments, answering SAT II questions led to positive testing effects: Students answered more final cued-recall questions correctly if the questions had been tested on the prior multiple-choice test. This effect was observed in both undergraduates and high school students, although the size of the effect differed across groups and depended on initial multiple-choice performance. Similarly, the effect was observed in all academic domains tested, but again the size of the effect depended on initial multiple-choice performance. The positive testing effect was stronger for domains and people associated with higher scores on the multiple-choice test. Multiple-choice scores tended to be better for biology, and testing benefited performance on cued-recall questions about biology the most. Undergraduates who scored highest on the initial multiple-choice test benefited more from taking the multiple-choice test than did the undergraduates with the worst scores. Similarly, on average, the undergraduates scored better on the SAT II questions than did high school students (as expected), and correspondingly the undergraduates showed the larger positive testing effect.

Answering SAT II questions also increased production of the multiple-choice distractors on the final test—a negative suggestion effect— but only in one case did the overall error rate increase after taking the test (see Figure 1). In essence, for undergraduates, answering SAT II questions replaced one type of error on the final test with a different error, one of the distractors from the SAT II. For undergraduates, the overall benefits of testing still outweighed the negative suggestion effect, even in the forced-responding condition of Experiment 2 and even for the subjects who scored in the bottom 25% on the multiple-choice test. In contrast, the high school students showed a larger negative suggestion effect than did the undergraduates, consistent with the finding that the high school students were much more likely to select distractors when answering the earlier multiple-choice questions. Even though testing reduced other wrong answers in high school students, this reduction was not large enough to offset the large increase in multiple-choice distractor intrusions, meaning that the overall effect of testing was null (in two domains) or negative (in two domains) for high school students. These data demonstrate the importance of assessing the consequences of testing for all types of students who may take the tests; focusing on high-achieving college students may lend an overly positive flavor to one's conclusions.

Two behaviors underlie the negative testing effect: selection of multiple-choice distractors on the initial test, and persistence of those selections over time. These two behaviors may sometimes be correlated, but they do not have to be, so we consider them separately here. First, students will be more likely to select multiple-choice distractors when they do not have the knowledge needed to separate the correct and incorrect alternatives. In our data set, this is likely why high school students selected more distractors than did college students. When free responding is allowed, selecting a distractor may also represent a failure of metacognition, in that the student fails to judge what she does not know and instead selects an incorrect alternative. In general, variables that increase multiple-choice distractor selection (such as less educated populations or forced-responding instructions) should increase the negative testing effect.

However, we want to be very clear that selecting a multiple-choice distractor does not guarantee a negative testing effect. Not all multiple-choice distractors that are selected on an initial test will persist to later tests. Across experiments and conditions, we observed persistence rates of 63% (Experiment 1), 61% (Experiment 2, free-responding condition), 39% (Experiment 2, forced-responding condition), and 43% (Experiment 3). It seems likely that distractors are more likely to persist when students are confident in their choices. Confidently held beliefs are likely stronger memories (e.g., Butterfield & Metcalfe, 2006) and also are likely to come from familiar domains that provide structure with which to associate the incorrect alternative. When a subject is forced to guess (e.g., in the forced-responding condition of Experiment 2), or simply does not know much about the tested topics (e.g., the high school students in Experiment 3), persistence of selected distractors will be lower. However, even though group (high school) and forced-responding instructions led to lower persistence rates, negative testing effects increased because of the large number of distractors selected on the initial multiple-choice test.

In all three experiments presented here, the overwhelming proportion of distractors produced on the final test had been selected on the original multiple-choice test. One way to reduce later intrusions of multiple-choice distractors is to administer feedback shortly after the initial test (Butler & Roediger, 2008). SAT practice books allow students to score their exams; a key is provided that links each question number to the letter of the correct alternative (e.g., 1 – A, 2 – C, etc.). This means that for the feedback to provide any information beyond "correct" or "incorrect" (which is not the best type of feedback; Pashler, Cepeda, Wixted, & Rohrer, 2005), the student must look back to the original question. We recommend stressing to students the importance of not only scoring their practice tests but also rereading the question and correct answer. This recommendation fits nicely with the literature on formative assessment, whereby testing paired with feedback is considered a learning opportunity (e.g., Black & William, 1998; Nicol & MacFarlane-Dick, 2006; Sadler, 1989).

The data from Experiment 2 (with undergraduates) suggest a second method for reducing the negative suggestion effect: Instructions that discouraged guessing on the initial test (the free-responding condition) led to a smaller negative suggestion effect on the final test. However, whereas the undergraduates benefited from being able to skip questions, the data from the high school students suggest that this instruction may not benefit all students. To be effective, skipping requires one to be able to judge what one does versus does not know, which requires metacognitive sophistication. This may also be a skill that requires practice; the high school students tested were from the Midwest, where the ACT (which does not penalize for wrong answers) is more common than the SAT (although they were familiar with the PSAT, which does penalize wrong answers).

The current research is only a first step in examining effects that answering standardized questions have on later test performance and suggests many interesting directions for future research. Our experiments examined performance on a test given shortly after students answered SAT II questions, so one open question is whether the effects of testing will persist over longer delays. In our work with other materials, we found delay reduced both the positive testing effect and the negative suggestion effect (Fazio et al., 2008). On the other hand, positive testing effects have also been shown over much longer delays (e.g., Spitzer, 1939), so the issue is still open. Finally, delay until the final test is, of course, just one of the differences between our participants' experiences and those of students taking real standardized tests (which are longer, involve more stress, and use questions that have never been in the public domain, to name just a few of the differences). The data that exist thus far suggest that the same kind of positive testing effects that are observed in the laboratory can also be observed in real classrooms (e.g., Bangert-Drowns, Kulik, & Kulik, 1991; Leeming, 2002; McDaniel, Anderson, Derbish, & Morrisette, 2007), suggesting that it would be fruitful for future research to examine testing effects from actual standardized tests.

In sum, we have shown that answering SAT II questions changes performance on a later test, and that these effects may vary depending on student ability. The scope and limits of such testing effects, and whether they will be observed with other tests and with other samples of students more heterogeneous in their abilities, must await future research. However, our results should stimulate research aimed at understanding how tests may be learning tools in addition to assessment tools.

References

Bacon, F. T. (1979). Credibility of repeated statements: Memory for trivia. *Journal of Experimental Psychology: Human Learning and Memory, 5,* 241–252.

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. (1991). Effects of frequent classroom testing. *Journal of Educational Research, 85,* 89–99.

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* (pp. 123–144). New York: Wiley.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5,* 7–74.

Bloom, B. S. (1956). *Taxonomy of educational objectives: Handbook I. Cognitive domain*. New York: Longman.

Butler, A. C., Marsh, E. J., Goode, M. K., & Roediger, H. L., III. (2006). When additional multiple-choice distractors aid versus hinder later memory. *Applied Cognitive Psychology, 20,* 941–956.

Butler, A. C., & Roediger, H. L., III. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36,* 604–616.

Butterfield, B., & Metcalfe, J. (2006). The correction of errors committed with high confidence. *Metacognition and Learning, 1,* 69–84.

College Board. (2002). *Real SAT II: Subject tests* (2nd ed.). New York: Author.

Fazio, L. K., Agarwal, P. K., Marsh, E. J., & Roediger, H. L., III. (2008.) *Consequences of multiple-choice testing persist over one week*. Manuscript submitted for publication.

Gates, A. (1917). Recitation as a factor in memorizing. *Archives of Psychology, 6,* No. 40.

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. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and 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 and Verbal Behavior, 10,* 562–567.

Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008). *Validity of the SAT for predicting first-year college grade point average* (College Board Research Report No. 2008–5). New York: College Board.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103,* 490–517.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin, 127,* 162–181.

Leeming, F. C. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology, 29,* 210–212.

Marsh, E. J., Roediger, H. L., III, Bjork, R. A., & Bjork, E. L. (2007). Memorial consequences of multiple-choice testing. *Psychonomic Bulletin and Review, 14,* 194–199.

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.

McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology, 16,* 192–201.

McDaniel, M. A., & Masson, M. E. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 371–385.

Nicol, D. J., & MacFarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Higher Education, 31,* 199–218.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 3–8.

Rees, P. J. (1986). Do medical students learn from multiple-choice examinations? *Medical Education, 20,* 123–125.

Remmers, H. H., & Remmers, E. M. (1926). The negative suggestion effect on true-false examination questions. *Journal of Educational Psychology, 17,* 52–56.

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, and Cognition, 31,* 1155–1159.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18,* 119–144.

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

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 Psychology, 86,* 357–362.

Received February 20, 2008
Revision received October 20, 2008
Accepted October 24, 2008 ■