Journal of Experimental Psychology: Applied
2011, Vol. 17, No. 4, 382–395

1076-898X/11/$12.00 DOI: 10.1037/a0026252

Test-Enhanced Learning in the Classroom: Long-Term Improvements From Quizzing

Henry L. Roediger III, Pooja K. Agarwal, Mark A. McDaniel, and Kathleen B. McDermott
Washington University in St. Louis

Three experiments examined whether quizzing promotes learning and retention of material from a social studies course with sixth grade students from a suburban middle school. The material used in the experiments was the course material students were to learn and some of the dependent measures were the actual tests on which students received grades. In within-subject designs, students received three low-stakes multiple-choice quizzes in Experiments 1 and 2 and performance on quizzed items was compared to that on items that were presented twice (Experiment 2) or items that were not presented on the initial quizzes (Experiments 1 and 2). We found that students' performance on both chapter exams and semester exams improved following quizzing relative to either not being quizzed or relative to the twice-presented items. In Experiment 3, students were given one multiple-choice quiz in class and encouraged to quiz themselves outside of class using a Web-based system. The assessment in this experiment was a short answer test in which students had to produce answers, but we also used multiple-choice tests. Once again, we found that quizzing of material produced a positive effect on chapter and semester exams. These results show the robustness of retrieval practice via testing as a learning mechanism in a classroom setting using the subject matter of the course and (in most cases) the tests on which students received grades as the dependent measures. Our results add to a growing body of evidence that retrieval practice in the classroom can boost academic performance.

Keywords: test-enhanced learning, testing effect, retrieval practice, classroom learning

A critical goal of classroom education is learning and retention of cognitive skills (reading, arithmetic, solving problems) and a huge number of facts (in virtually all subjects, but particularly in subjects such as science, history, social studies, and the like). Part of becoming knowledgeable in any subject is mastering the large body of facts that represent its subject matter. Because research on learning and memory is aimed at fact learning, one might think that this research should be relevant to education. We think it is, and many books attest to the value of various strategies to improving fact learning in the classroom (e.g., Mayer, 2008; Mayer, 2010; Willingham, 2009).

This article was published Online First November 14, 2011.

Henry L. Roediger III, Pooja K. Agarwal, Mark A. McDaniel, and Kathleen B. McDermott, Department of Psychology, Washington University in St. Louis.

This research was supported by 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 Institute or the U.S. Department of Education. We are grateful to the Columbia Community Unit School District 4, superintendents Leo Sherman, Jack Turner, and Ed Settles, Columbia Middle School principal Roger Chamberlain, social studies teacher Patrice Bain, and all of the 2006-2008 sixth grade students and parents. We also thank Lindsay Brockmeier and Kristy Duprey for their help preparing materials and testing students, and Jane McConnell, Kari Farmer, and Jeff Foster for their assistance throughout the project.

Correspondence concerning this article should be addressed to Henry L. Roediger, III, Department of Psychology–Box 1125, Washington University, One Brookings Drive, St. Louis, MO 63105-4899. E-mail: roediger@wustl.edu

The traditional approach to enhancing learning and retention, both in laboratory studies and by their extension to the classroom, is to change study strategies. In the lab, researchers emphasize organizational schemes (e.g., Mandler, 1967; Tulving, 1962; Tulving, 1968), mental imagery (e.g., Paivio, 1969), or the types of processing provided (Craik & Tulving, 1975). With text materials, researchers have correspondingly emphasized the importance of organizational structures and text coherence (e.g., Kintsch, 1998), relational or item-specific processing during reading (Hunt & McDaniel, 1993), and similar tactics that focus on learning from texts while studying. When surveys have asked university students about their study strategies in preparing for tests, the great majority report reading a text, underlining or highlighting it, and then reviewing the highlighted parts (e.g., Karpicke, Butler, & Roediger, 2009; Kornell & Bjork, 2009). Some form of repeated reading forms the primary study strategy for most university students.

Repeated reading makes students fluent in processing material and increases estimates that they know it well (and can, therefore, cease studying; e.g., Karpicke, 2009; Roediger & Karpicke, 2006b). However, the students' challenge on tests (especially short answer and essay tests) is not merely to process information fluently, but to recall the information when given either specific cues (short answer tests) or often very general cues (as on essay tests). Because these tests require relatively effortful retrieval from memory, one can wonder if repeated reading (which permits gains in fluency but does not permit retrieval practice) is the most effective study strategy (e.g., Callender & McDaniel, 2009 provide evidence that repeated reading of textbook chapters is ineffective on a later test relative to a single reading). Applying the principle of transfer appropriate processing (Bransford, Franks, Morris, &

Stein, 1979), one might expect that the best way to foster good performance on a test requiring active retrieval would be to have students practice such retrieval. That is precisely the tactic taken in the current research, although we show that active retrieval can lead to gains even on multiple-choice recognition tests presumed to require less effortful retrieval than short answer or essay tests.

A century of research has been devoted to study of the testing effect—the fact that active retrieval produces better retention than passive rereading—although the continuity and progress in understanding the effect may best be described as erratic. Abbott (1909) published the first study on the topic, and shortly after her work, Gates (1917) and Jones (1923−1924) also showed that taking a test can be an effective learning tool. Studies of the testing effect have different groups of subjects study the same material. One group takes one or more tests after study, whereas the other group either has no further dealings with the material or (in another type of control) rereads the material the same number of times that the testing group is tested. All subjects are tested on a final retention test some time later. The great body of research on this topic (see Roediger & Karpicke, 2006a and Roediger, Agarwal, Kang, & Marsh, 2010 for reviews) shows that taking a test confers a much greater benefit than not taking a test (e.g., Wheeler & Roediger, 1992); further, testing usually provides a benefit even relative to repeated restudy of the material, and this is especially true on delayed tests (e.g., Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008; Hogan & Kintsch, 1971; Roediger & Karpicke, 2006b; Wheeler, Ewers & Buonanno, 2003). Interestingly, the power of testing seems to increase with the number of tests taken and also when tests are followed by feedback (e.g., Roediger & Karpicke, 2006b).

Given the power of retrieval practice during testing in benefiting retention, the use of this strategy in enhancing performance in educational settings seems natural. After all, tests are given as a part of classroom instruction, although almost always for purposes of assessing students' knowledge and assigning grades. Although educators sometimes decry the emphasis on testing in schools, usually they have in mind high-stakes standardized tests, which determine students' placement, graduation, or college admission. Our use of testing is to provide retrieval practice on information via low stakes quizzes that count little or nothing toward a student's grade in the course. Rather, the quiz serves two other important functions: first, testing (especially with feedback) enhances learning and retention of the material, and second, the metacognitive use of tests lets students inform themselves about what they know and do not know so they can concentrate future study efforts on the information that they do not know. In fact, in those relatively rare cases when students report using self-testing as a study strategy, they usually cite the second reason for doing so and not the first (Karpicke et al., 2009; Kornell & Bjork, 2009). In addition, repeated studying typically inflates students' judgments or predictions of learning; students have more accurate metacognitive judgments following testing (Agarwal et al., 2008).

McDaniel, Roediger, and McDermott (2007) outlined this approach to enhancing educational performance that we call test-enhanced learning. However, relatively few studies have shown that testing can work in an actual educational setting. Many studies have been done in classrooms, but almost always the material used is extraneous to the course. In a powerful series of experiments, Gates (1917) showed that testing improved retention of nonsense words and poetry, among other materials. Spitzer (1939) showed that testing improved performance in sixth grade students, but the material he used (passages on peanuts and bamboo) was not part of their regular curriculum. Recently, McDaniel, Anderson, Derbish, and Morrissette (2007) used testing in a systematic experiment in a college course. Students in a web-based course on brain and behavior either took quizzes or reread critical facts and then these facts were later tested. The results showed that prior quizzing produced higher scores than did rereading. One problem with the study is that students were assessed only on peripheral facts rather than on central material that was tested in the course. This procedure was followed because of concerns by the university's IRB about testing in a way that would potentially affect students' grades. Baseline levels of performance were quite low (44%), probably because of the tangential nature of the material selected for the experiment. Many other studies have also shown that testing works in classroom settings but the material, quizzes, and/or criterial tests used in these studies were not integral to the course (e.g., Carpenter, Pashler, & Cepeda, 2009; Duchastel & Nungester, 1982; Sones & Stroud, 1940; Swenson & Kulhavy, 1974; but see Glass, 2009; McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011, for an exception).

Before quizzing can be suggested as a method to enhance educational performance, it is necessary to show that it works in the classroom at many different levels in the educational system. In the current article, we report results from three experiments in which we used the actual content of the course for our experiments and the chapter tests and semester exams in the course were our dependent measures. We conducted true experiments to examine retrieval practice (quizzing) in within-student designs in a sixth grade social studies classroom. With full cooperation of the teacher and the school administrators, we included six classrooms of students who were all taking the same course. The use of multiple classrooms permitted us to rotate material through various conditions across the classes, all the while using procedures that kept the teacher unaware of which material was assigned to which condition in each class. We used the texts assigned in the classroom and we also used the set of materials (including tests) the teacher had already developed. Because we sought to maintain normal classroom practices, we used multiple-choice tests that the teacher had developed as quizzes, and we used both multiple-choice and free recall exams (explained below) as the final criterial tests. The questions for the quizzes repeated the same material as on the final criterial test because of constraints placed on the research by using actual course materials. However, in Experiment 3 we used short answer questions (rather than multiple-choice) for the criterial test. Each experiment reports research conducted over the better part of a semester (four chapters of material) with a research assistant in the classroom every day to administer the quizzes. Thus, the research reported here represents data collected over 1.5 years (three semesters) in a sixth grade social studies classroom. The data we report came from the tests and exams on which students received grades.

The critical question addressed in our experiments was whether quizzes would enhance final retention relative to control conditions that involved no quizzing of material or that involved rereading the material. In Experiment 1 we examined whether a basic testing effect would occur in the classroom using the minimal necessary conditions involving tested versus nontested items.

Experiment 1

Method

Participants. One hundred forty-two sixth grade social studies students from a public middle school located in a Midwestern suburban, middle-class community participated in this study. Parents were informed of the study and written assent from each student was obtained in accordance with guidelines of the Human Research Protection Office.

Materials and design. In sixth grade social studies at this school, students learn about cultures and their history from around the world. We used material from four chapters in the assigned social studies textbook (Ancient Egypt, Mesopotamia, India, and China), presented in this order as determined by the classroom teacher. On initial classroom quizzes (pretests before the teacher's lesson, posttests after the teacher's lesson, and review tests a few days later), half of the target facts from each chapter were tested in a multiple-choice format (tested condition) and half of the facts were not tested (nontested condition), using a within-students design (that is, whether or not material was quizzed was manipulated within-students). The number of target facts varied across chapters (32, 24, 28, and 20 items, respectively), half of the target facts from each chapter were randomly assigned to the two conditions, and each of the six classroom sections received a different random selection of items. The total number of items in this experiment (across all four chapters) was 104, or 52 items per condition.

For example, a multiple-choice fact included:

What is Pharaoh Tutankhamun best known for?

(a) The way he ruled his kingdom

(b) Living to an old age

(c) The belongings found in his tomb

(d) His trading routes with other kingdoms

For initial quizzes (pre-, post-, and review), a research assistant administered the classroom quizzes orally and visually using a clicker response system (Ward, 2007). After responding to each multiple-choice question, students were provided with immediate feedback in the form of a green checkmark next to the correct answer while the experimenter read aloud the question stem and correct answer. Questions on the initial quizzes were presented in the order in which they appeared in the chapter. The four multiple-choice alternatives were presented in a different random order for each pre-, post-, and review test.

To measure retention, the classroom teacher administered chapter exams in paper and pencil format. The chapter exam generally occurred two days after the review quiz. The first part of the chapter exam consisted of a free recall exam in which students were asked to write down everything they remembered from the chapter. Following the free recall exam, students completed a multiple-choice exam comprised of all tested and nontested items. Multiple-choice questions on the chapter exams were the same as those on the initial classroom quizzes, presented in the same order for each classroom section. The four multiple-choice alternatives were reordered randomly. Students received delayed feedback from the classroom teacher approximately 2 days after the chapter exam.

The part of the experiment just described was carried out during the fall semester (between the beginning of school and the winter break). Besides the chapter exams, students also completed multiple-choice exams at the end of the semester and at the end of the academic year. These delayed exams were administered via the clicker response system and were not counted as part of students' grades. Questions were presented in the order in which the chapters appeared in the textbook, and questions for each chapter were presented in a different random order for each classroom section. For example, items from chapter 4 were presented in random order followed by items from chapter 5 presented in random order, and so forth. The end-of-the-semester exam was composed of eight target facts (four items per condition) from each of the four chapters, and the end-of-the-year exam was composed of four target facts (two items per condition) from each of the four chapters. All facts were tested at least once on the chapter exam, yet items on the end-of-the-year exam were not presented on the end-of-the-semester exam (to avoid repeated testing effects on these long-delayed tests). However, due to the small number of items on the end-of-the-year exam, results were highly variable among students and thus, are not reported. In other experiments with more observations at the end of the year, we have obtained significant effects of initial testing (McDaniel, Agarwal, et al., 2011).

Procedure. Students were tested in classroom sections ranging from 21 to 27 students each, using a within-subjects design. Before the teacher's lesson, students took a pretest over the items that were later to be tested. The teacher stepped outside the room while the project's research assistant administered the pretest, so the teacher was not present and, therefore, did not know which target facts were quizzed or not quizzed. Immediately following the pretest, the teacher taught the lesson for the day, which covered target facts, both tested and nontested facts. Immediately after the lesson, students took a posttest over tested items. Approximately two days later, students took a review test over tested items. The teacher was present for posttests and review tests, but because each of six classroom sections received a different random assignment of items per condition, it is unlikely that the teacher could keep track of which target facts were quizzed or not quizzed.

Retention was measured at the end of the chapter (*M* length of time for covering chapter content = 13.25 days) via free recall and multiple-choice exams. For the free recall exam, students were given a blank sheet of paper and asked to write down everything they could remember from the chapter. A list of all facts in the chapters was constructed, and the data were scored for the number of facts recalled. These were converted to proportions for analysis. Long-term retention was also measured on unanticipated multiple-choice end-of-the-semester and end-of-the-year exams; students were not informed of these exams in advance. Depending on when the chapter exam was given, the semester exam occurred approximately 1–2 months later.

Results and Discussion

Preliminary considerations. Twenty-three students who qualified for special education or gifted programs were excluded from our analyses. The special education students received considerable further study outside of the classroom (including some testing), and the gifted students were at or near ceiling on the quizzes and chapter tests even in the control condition. Furthermore, 82 students who were not present for all initial quizzes,

chapter exams, and delayed exams were also excluded to ensure the integrity of our testing schedule. Therefore, data from 36 students who met the criteria of being present for all quizzes and exams are reported below. However, despite these exclusions, the general pattern of results remained the same when data from the 82 students who were absent for part of the manipulation were included, confirmed by additional analyses (which are not reported for brevity). These data for all 118 students (excluding only gifted and special education students) are shown in Table A-1 of the Appendix (collapsed across chapters). It was confirmed via additional analyses that for this and for the other two experiments described below (whose data are shown in Tables A-2 and A-3) that analyzing data of only students who completed the entire design does not bias the findings.

Because there were unequal numbers of items in the chapters, all means were weighted proportionally. The variable of "chapter" serves mainly to show replicability of our basic results across materials. Of course, the "chapter" variable is confounded with order, because the teacher assigned the chapters in the same order (the order in which they appeared in the book) across all six of her classes. However, as indicated below, most of our results generalize well across chapters, with some differences in magnitude of effects in the various chapters creating interactions. All results deemed significant in these experiments exceeded an alpha level of .05 unless otherwise noted.

Initial quiz performance. Initial quiz performance as a function of chapter and type of test is shown in Table 1 and reveals general improvement across the quizzes. A 3 (quiz type: pretest, posttest, review) × 4 (chapter) repeated measures analysis of variance (ANOVA) confirmed a significant increase from the pretests (41%) to the posttests (93%) and review tests (93%), $F(2, 70) = 1188.28$, $\eta_p^2 = .97$. Pairwise comparisons indicated that posttest performance and review test performance were significantly greater than pretest performance, $t(35) = 39.91$, $d = 7.00$ and $t(35) = 34.92$, $d = 6.97$, respectively, although posttest and review test performance were near ceiling and did not significantly differ, $t < 1$. These results demonstrate substantial student learning from the teacher's lesson between pre- and posttests given with feedback, and little forgetting between post- and review tests (with the exception of the chapter on ancient China). Quiz performance also tended to differ depending on the chapter, $F(3, 105) = 209.83$, $\eta_p^2 = .86$; this effect probably indicates that chapters and/or test items on those chapters varied in difficulty. Posttest performance was generally greater than pretest performance, and review test performance was greater than posttest performance, except for an unusual drop in performance on the review test for the chapter on ancient China chapter, which created an interaction, $F(6, 210) = 16.28$, $\eta_p^2 = .32$, but which is difficult to interpret. The main point is that performance dramatically increased across the initial quizzes.

Chapter exam performance: Free recall. Free recall performance as a function of chapter and learning condition is shown in the top two rows of Table 2. For the free recall exam, students were given the topic (e.g., Ancient India) and simply asked to recall all the facts they had learned about that topic. The data show the mean proportion of idea units in the chapter material that were recalled, and they reveal a strong testing effect for three of the four chapters (with the chapter on China again representing the exception). Previously quizzed items were recalled better than those presented only in the lectures and the readings for three chapters. A 2 (learning condition: tested, nontested) × 4 (chapter) repeated measures ANOVA produced a main effect of learning condition; performance on the free recall exam was significantly greater for tested items (30%) than for nontested items (20%), $F(1, 35) = 63.22$, $\eta_p^2 = .64$. Performance varied across the chapters, $F(3, 105) = 24.23$, $\eta_p^2 = .41$, and the benefit from quizzing also varied across the four chapters of material, $F(3, 105) = 9.04$, $\eta_p^2 = .21$ for the interaction. We have no ready explanation for why the testing effect differed across the four chapters of material and did not occur for the chapter on ancient China. The answer may have more to do with the particular facts chosen for some chapters being more likely to produce testing effects than others, but as discussed below all four chapters did show testing effects on the multiple-choice exam. The important point is that testing effects were obtained on three of the four chapters for the free recall exam, $ps < .05$.

Chapter exam performance: Multiple-choice. Multiple-choice performance as a function of chapter and learning condition is shown in the third and fourth rows of Table 2, and again a testing effect is apparent. A 2 (learning condition: tested, nontested) × 4 (chapter) repeated measures ANOVA confirmed that performance was greater for tested items (94%) than nontested items (81%), $F(1, 35) = 95.66$, $\eta_p^2 = .73$, indicating the robust effects of quizzing on retention in a classroom setting. The main effect of chapter, $F(3, 105) = 279.42$, $\eta_p^2 = .89$, simply shows that material (or test items) for some chapters were easier than for others. The interaction between learning condition and chapter material was also significant, $F(3, 105) = 2.71$, $p = .049$, $\eta_p^2 = .07$.

The multiple-choice chapter exams were the tests on which students' grades were largely based, so performance on these tests is of special interest. The control (nontested) items led to 81% correct, which the teacher told us is her usual level of performance and represents roughly a B− grade. Given this baseline, quizzing could only potentially show a 19% improvement. The quizzed items showed a 13% increase, so quizzing enhanced performance

Table 1
Initial Quiz Performance as a Function of Chapter and Type of Quiz in Experiment 1

	Ancient Egypt	Ancient Mesopotamia	Ancient India	Ancient China	Mean
Pretest	.42 (.15)	.40 (.14)	.42 (.15)	.39 (.12)	.41 (.09)
Posttest	.91 (.10)	.96 (.06)	.92 (.09)	.95 (.07)	.93 (.05)
Review test	.95 (.06)	.94 (.07)	.94 (.08)	.86 (.14)	.93 (.05)

Note. Overall means have been weighted according to the number of items per chapter. Standard deviations are shown in parentheses.

Table 2
Chapter and Semester Exam Performance as a Function of Chapter, Test Format, and Learning Condition in Experiment 1

	Ancient Egypt	Ancient Mesopotamia	Ancient India	Ancient China	Mean
Chapter Exam: Free Recall					
Tested	.38 (.16)	.33 (.19)	.20 (.11)	.27 (.18)	.30 (.10)
Nontested	.24 (.15)	.14 (.12)	.14 (.14)	.30 (.17)	.20 (.09)
Chapter Exam: Multiple-Choice					
Tested	.95 (.07)	.94 (.09)	.92 (.08)	.94 (.10)	.94 (.05)
Nontested	.86 (.11)	.77 (.14)	.78 (.16)	.84 (.11)	.81 (.09)
End-of-the-Semester					
Tested	.76 (.26)	.83 (.22)	.70 (.19)	.89 (.14)	.79 (.14)
Nontested	.74 (.26)	.68 (.24)	.57 (.28)	.69 (.23)	.67 (.19)

Note. Overall means have been weighted according to the number of items per chapter. Standard deviations are shown in parentheses.

by 68% of the possible gain (.13/.19 × 100). If all items (rather than only a subset) had been quizzed, students' grades would have been lifted from a B− to an A.

Semester exam performance. Performance on the multiple-choice end-of-the-semester exam is displayed in the bottom two rows of Table 2. A 2 × 4 ANOVA was computed, and on the end-of-the-semester exam, performance on tested items (79%) was greater than nontested items (67%), $F(1, 35) = 28.73$, $\eta_p^2 = .45$. Performance again significantly varied across chapters, $F(3, 105) = 30.23$, $\eta_p^2 = .46$, although the interaction between chapter and learning condition was not significant, $F(3, 105) = 2.03$, $p = .11$. Although tested performance was better than nontested for all chapters, the effect was significant for only the three chapters with the larger effects ($ps < .05$), but not for the chapter on ancient Egypt ($p = .77$).

In sum, substantial testing effects were obtained on free recall and multiple-choice chapter exams. The testing effect persisted until the end of the semester on a multiple-choice exam, which occurred at a relatively long delay (1–2 months) for some chapters. Although the multiple-choice quizzes used the same items repeatedly (with random placement of the target among three lures), other studies have shown that quizzing enhances retention on restated versions of the target questions (McDaniel, Thomas, Agarwal, Roediger, & McDermott, 2011) and on items requiring transfer of knowledge to new situations (Butler, 2010). Of course, we also showed a quizzing effect on production of target facts in free recall (see Table 2), as well.

Experiment 2

The results of Experiment 1 showed relatively consistent testing effects on exams given at the end of chapters and at the end of the semester. The testing effect was robust across subject matters, especially on chapter exams. However, a critic could complain that Experiment 1 simply showed that reviewing material is advantageous. That is, because the quizzes occurred with feedback, the testing effect may simply have been due to repeated studying of the material being more beneficial than less studying (e.g., Thompson, Wenger, & Bartling, 1978). Certainly reexposure may form part of the testing effect, because feedback given after testing produces test-potentiated learning (that is, people learn more from a study presentation if it is preceded by a test; e.g., Izawa, 1971). Many researchers have compared repeated study conditions to repeated test conditions and have shown that repeated testing is usually superior to repeating studying, especially on delayed tests that involve recall (see Roediger & Karpicke, 2006a, pp. 197–198 for a review). However, some experiments comparing recognition tests to rereading control conditions on a later criterial test have not obtained an advantage of testing to restudying (see Butler & Roediger, 2007; Kang, McDermott & Roediger, 2007; McDaniel, Anderson, Derbish, & Morrisette, 2007). These prior experiments were mostly laboratory experiments over shorter time intervals than used in the classroom experiments of Experiment 1 here. In Experiment 2 we asked whether repeated quizzing would permit greater gains on a chapter exam relative to repeated studying, thus, confirming that quizzing has benefits over and above restudying. We also included the condition in which some materials received neither repeated reading nor repeated quizzing for comparison.

Method

Participants. The same one hundred forty-two students from Experiment 1, in addition to one new student, participated in Experiment 2. Experiment 2 was conducted during the spring semester of the same academic year.

Materials and design. We used material from three chapters in the assigned social studies textbook (Ancient Rome, the Middle Ages, and Africa), presented in this order as determined by the classroom teacher. Critically, a read control condition was added in Experiment 2 for comparison with tested and nontested conditions used in Experiment 1. For example, a corresponding read fact to the multiple-choice fact described in Experiment 1 would be presented on classroom quizzes as: "Pharaoh Tutankhamun is best known for the belongings found in his tomb."

On initial classroom quizzes (pre-, post-, and review tests), approximately a third of the target facts from each chapter lesson were tested in a multiple-choice format, a third of the target facts were presented for reading, and a third of the facts were not tested, again using a within-subjects design. Presentation of tested and read items on initial quizzes was mixed (i.e., not blocked by condition). The number of target facts varied across chapters (32, 25, and 28 items, respectively), they were randomly assigned to the three conditions, and each of the six classroom sections received a different random selection of items. The total number of items in this experiment was 85: 32 items in the tested condition, 28 items in the read condition, and 25 items in the nontested condition.

For initial quizzes (pre-, post-, and review), an experimenter administered the classroom quizzes orally and visually using a

clicker response system (Ward, 2007). After responding to each multiple-choice item, students were provided with immediate feedback in the form of a green checkmark next to the correct answer while the experimenter read aloud the question stem and correct answer. For each read item, the experimenter read aloud the answer statement while students followed along. Items on the initial quizzes were presented in the order in which they appeared in the chapter. The four multiple-choice alternatives for tested items were presented in a different random order for each pre-, post-, and review test.

Paper and pencil chapter exams were composed of all target facts (those initially tested, read, and nontested) in multiple-choice format; free recall tests were not used. Multiple-choice questions on the chapter exams were the same as those on the initial classroom quizzes, presented in the same order for each classroom section. The four multiple-choice alternatives were reordered randomly. Students received delayed feedback from the classroom teacher approximately 2 days after the chapter exam.

In addition, students completed an unanticipated multiple-choice exam at the end of the semester (approx. 3–5 months later), which was composed of six target facts from each of the three chapters (two items per condition) and was administered via the clicker response system. Questions were presented in the order in which the chapters appeared in the textbook, and questions for each chapter were presented in a different random order for each classroom section. Because this experiment was conducted during the spring semester, the semester exam occurred at the end of the academic year.

Procedure. Procedures were similar to those used in Experiment 1. Before the teacher's lesson, students took a pretest that included quizzed and read items. The teacher was not present for the pretest and did not know which target facts were tested, read, or nontested; further, the items assigned to each condition differed across classes. Following the pretest, the teacher taught the lesson for the day, which covered all target facts. Immediately after the lesson, students took a posttest that included tested and read items. To reiterate, students were encouraged to attend to the read items and the experimenter read them aloud while the student read them silently. Approximately 2 days later, students took a review test including tested and read items. Retention was measured at the end of the chapter (*M* length of time for covering chapter content = 11.33 days) on multiple-choice exams composed of all target facts, as well as on an incidental multiple-choice end-of-the-semester exam composed of all target facts.

Results and Discussion

Twenty-three students who qualified for special education or gifted programs were excluded from our analyses. Furthermore, 56 students who were not present for all initial quizzes, chapter exams, and delayed exams were also excluded to ensure the integrity of our testing schedule. Therefore, data from 63 students are reported below, but (as in Experiment 1) the pattern of results, confirmed by additional analyses, remained the same when the 56 absent students were included (see Appendix, Table A-2). As in analysis of Experiment 1, means have been weighted proportionally.

Initial quiz performance. Initial quiz performance as a function of chapter and type of test is shown in Table 3, where a sharp increase in performance can be seen from the pretest to the post- and review tests. A 3 (quiz type: pretest, posttest, review test) × 3 (chapter) repeated measures ANOVA confirmed a significant increase from the pretests (41%) to the posttests (89%) and review tests (87%), $F(2, 124) = 840.42$, $\eta_p^2 = .93$. Pairwise comparisons confirmed that posttest and review test performance were significantly greater than pretest performance, $t(62) = 33.21$, $d = 4.70$ and $t(62) = 32.95$, $d = 4.65$, respectively, and posttest performance was significantly greater than review test performance, $t(62) = 2.28$, $d = .33$, even though this was only a 2% difference. These results demonstrate substantial student learning from the teacher's lesson between pre- and posttests, and little forgetting between post- and review tests. Initial quiz performance also differed across the three chapters, $F(2, 124) = 141.73$, $\eta_p^2 = .70$, with performance lowest for the Africa chapter. For all three chapters, posttest performance was greater than pretest performance, but review test performance was less than posttest performance only for the Africa chapter, indicated by a significant chapter by quiz type interaction, $F(4, 248) = 7.49$, $\eta_p^2 = .11$. Performance was not as great on the post- and review tests as for the chapters in Experiment 1, because either the subject matter or test items were more difficult.

Chapter exam performance. Multiple-choice performance as a function of chapter and learning condition is shown in the top three rows of Table 4, where a testing effect is again apparent. A 3 (tested, read, nontested) × 3 (chapter) ANOVA showed a main effect of testing: performance was greater for tested items (91%) than read (83%) and nontested items (81%), $F(2, 124) = 33.82$, $\eta_p^2 = .35$. Pairwise comparisons confirmed a significant testing effect (tested greater than nontested), $t(62) = 7.60$, $d = .98$, as well as a significant testing benefit relative to read items, $t(62) = 6.61$, $d = .83$, indicating the robust effects of quizzing on retention over and above reading statements in a classroom setting. This outcome confirms in a classroom study what many experiments have determined in the lab: testing benefits later retention more than restudying (e.g., Carrier & Pashler, 1992). Moreover, performance did not differ significantly between the read and nontested items, $t(62) = 1.64$, demonstrating that classroom reading did not signif-

Table 3
Initial Quiz Performance as a Function of Chapter and Type of Quiz in Experiment 2

	Ancient Rome	Middle Ages	Africa	Mean
Pretest	.45 (.20)	.46 (.17)	.34 (.19)	.41 (.12)
Posttest	.91 (.08)	.93 (.08)	.84 (.19)	.89 (.08)
Review test	.92 (.09)	.95 (.06)	.73 (.14)	.87 (.07)

Note. Overall means have been weighted according to the number of items per chapter. Standard deviations are shown in parentheses.

Table 4
Chapter and Semester Exam Performance as a Function of Chapter, Test Format, and Learning Condition in Experiment 2

	Ancient Rome	Middle Ages	Africa	Mean
Chapter Exam				
Tested	.94 (.08)	.96 (.08)	.82 (.14)	.91 (.07)
Read	.84 (.14)	.88 (.12)	.78 (.18)	.83 (.11)
Nontested	.89 (.12)	.80 (.22)	.73 (.19)	.81 (.12)
End-of-the-Semester				
Tested	.57 (.35)	.70 (.28)	.52 (.37)	.59 (.22)
Read	.53 (.37)	.56 (.38)	.51 (.37)	.53 (.23)
Nontested	.50 (.35)	.60 (.32)	.56 (.36)	.54 (.17)

Note. Overall means have been weighted according to the number of items per chapter. Standard deviations are shown in parentheses.

icantly increase retention of the target facts even though students read the items three times with spaced presentation. This outcome, too, has parallels with laboratory studies, even those using text materials (e.g., Callender & McDaniel, 2009). Repeated studying often has little effect on delayed tests (Karpicke & Roediger, 2008; Zaromb & Roediger, 2010). However, because we used spaced presentations of material, it does seem surprising that repeated studying had so little effect (Rawson & Kintsch, 2005). Chapter exam performance differed across the chapters, with the lowest performance for the Africa chapter, $F(2, 124) = 188.35$, $\eta_p^2 = .75$. The interaction between condition and chapter was also significant, $F(4, 248) = 6.06$, $\eta_p^2 = .09$, due to differences between performance for read and nontested items across chapters.

The important point is that quizzing led to an advantage over and above repeated reading of the material. A reviewer suggested that because we used a mixed format in which some items were presented as questions to be answered and some as facts to be studied, that perhaps students spent the time during the rereading periods to rehearse the question items. We cannot rule out this possibility, but we view it as remote. While students read the fact from the screen, the experimenter read it aloud and thus commanded students' attention. If we had used a blocked design in which tested and reread items were presented in separate groups, we think students' attention might have flagged from reading long lists of facts, which is why we chose the mixed design.

Semester exam performance. Performance on the multiple-choice end-of-the-semester exam is displayed in bottom three rows of Table 4 and showed only slender evidence that a testing effect occurred on this delayed test. On the end-of-the-semester exam, a 3 (tested, read, nontested) × 3 (chapter) ANOVA showed only the main effect of chapter to be significant, $F(2, 124) = 3.15$, $\eta_p^2 = .05$. Although the main effect of learning condition was not significant, performance for tested items in two of the three chapters was greater than for read items (e.g., a 5% difference for ancient Rome and a 12% difference for the Middle Ages). The testing effect for the Middle Ages was significant relevant to the read control, $t(62) = 2.21$, $d = .41$, and it was also significant relative to the not-tested control, $t(62) = 2.08$, $d = .34$. Still, evidence for a testing effect at the end of the semester, while suggestive, was weak.

In sum, substantial testing effects were obtained on chapter exams even over and above a rereading control condition. However, this benefit of quizzing largely disappeared on the end-of-the-semester exam that occurred 3–5 months later.

Experiment 3

The first two experiments have shown robust testing effects in classroom settings using actual course material. Further, quizzing proved superior both to a nontested control condition and to a repeated study condition. The gains from testing show some evidence of lasting at least until an end-of-the-semester exam (especially in Experiment 1). In Experiment 3, we asked two additional questions. First, if students were permitted to test themselves outside of class on the material, would the gains from self-testing add to the benefit of quizzes given in class? The teacher of the class had reported that students enjoyed quizzing themselves on an Internet-based system that permitted them to play games and earn points while correctly answering questions. Thus, we arranged for the tested (but not the nontested) material to appear on Quia Web (http://www.quia.com/web) for students to access outside of class (e.g., at home or in the school library).

The second question we asked in Experiment 3 was whether benefits from retrieval practice would extend to a final short answer exam. In the first two experiments, multiple-choice performance on the chapter exams was quite high. Thus, testing with short answer questions may permit us to examine performance at a lower point on the measurement scale. In addition, if we were able to obtain testing effects with short answer questions, we could rule out the possibility that the only feature students were learning from quizzes was how to pick out items from among the same four alternatives on a multiple-choice test. Of course, the fact that we obtained testing effects on the free recall test (somewhat similar to an essay test) in Experiment 1 makes this possibility unlikely a priori.

Method

Participants. One hundred thirty-two sixth grade social studies students from the same public middle school participated in this study. The experiment was conducted during the academic year following that in which Experiment 1 was conducted. Parents were informed of the study and written assent from each student was obtained in accordance with guidelines of the Human Research Protection Office.

Materials and design. During the fall semester, we used material from five chapters in the assigned social studies textbook (Ancient Egypt, Mesopotamia, India, China, and Greece), presented in this order as determined by the classroom teacher. The only initial classroom quizzes used in this experiment were pretests before the teacher's lesson. On pretests, half of the target facts from each chapter were tested in a multiple-choice format (tested condition) and half of the facts were not tested (nontested condition), using a within-subjects design. Target facts were randomly assigned to the two conditions and each of the six classroom sections received a different random selection. The number of target facts varied across chapters (ranging from 18–32 items per chapter), and the total number of items in this experiment was 125 items: 63 items in the tested condition and 62 items in the nontested condition.

For the initial pretest quizzes, an experimenter administered the classroom quizzes orally and visually using a clicker response system (Ward, 2007). After responding, students were provided with immediate feedback in the form of a green checkmark next to the correct answer while the experimenter read aloud the question stem and correct answer. Questions on the pretests were presented in the order in which they appeared in the chapter. The four multiple-choice alternatives were presented in a different random order for each pretest.

Additional self-quizzing by students of tested items via Quia Web (http://www.quia.com/web) was encouraged. Four kinds of games based on tested items uploaded by the experimenter were made available to students: matching, where students matched question stem cards (presented in one color) to answer cards (presented in another color); flashcards, where students read a question and turned over the online card to read the answer; concentration, where students could click on cards to uncover them and then try to match question stems and answers; and "columns," where students drew lines between question stems and answers placed in separate vertical columns. Because each class section had a different random selection of tested items, each section had a separate Web site with corresponding games. Unfortunately, due to limitations of the technology, we were unable to track when or how often individual students used the Web site.

To measure retention, the classroom teacher administered eight lesson exams in paper and pencil format over the course of five chapters, which occurred slightly more frequently than chapter exams in the previous experiments (M length of lesson = 4.13 days). Lesson exams comprised all tested and nontested items, half of which were in multiple-choice format, the other half in short answer format. A word bank, or a list of all key terms from the entire chapter, was included with each lesson exam to aid students on short answer questions. All questions on the lesson exams were presented in random order, except that multiple-choice questions were always followed by short answer questions on different content, and multiple-choice and short answer questions were the same for all students. At the end of each lesson exam, students were asked, "Did you use Quia outside of class to study for this lesson?" and if so, "How many times did you play games on Quia for this lesson?" Students received delayed feedback from the classroom teacher approximately 2 days after the lesson exam.

Students also completed unanticipated multiple-choice end-of-the-semester (approx. 1–3 months after chapter tests) and end-of-the-year exams (approx. 6–8 months after chapter tests), which were administered via the clicker response system. The end-of-the-semester exam was composed of eight target facts from each of the five chapters and the end-of-the-year exam comprised four different target facts from each of the five chapters. All facts were tested at least once on the lesson exams and a subset of these was retested on the end-of-semester and the end-of-the-year exams. Questions on delayed exams were presented in the order in which the chapters appeared in the textbook, and questions for each chapter were presented in a random order for each classroom section. Again, due to the limited number of items on the end-of-the-year exam, results were unreliable and thus, are not reported.

Procedure. Subjects were tested in classroom sections ranging from 15 to 28 students each, using a within-subjects design. Before the teacher's lesson, students took a pretest over tested items. The teacher was not present for the pretest and did not know which target facts were tested or nontested. Following the pretest, the teacher taught the lesson for the day, which covered all target facts, both tested and nontested facts.

Between the teacher's lesson and the lesson exam, students were encouraged to quiz themselves using the Web site in the classroom or school library during school hours, as well as from home. (All but two students reported having a computer in their homes that could be used to access Quia Web.) Lesson material was not available on the Web site until the teacher introduced the material in class; in other words, students could not work ahead, but they could access games from past lessons. Retention was measured at the end of the lesson on lesson exams comprised of multiple-choice and short answer questions over all target facts from the lesson at hand. For example, a short answer question corresponding to the multiple-choice question described in Experiment 1 would be presented on lesson exams as: "What is Pharaoh Tutankhamun best known for?" The experimenter, with help from the classroom teacher, scored all short answer questions as either correct or incorrect. Long-term retention was also measured on multiple-choice exams at the end of the semester and end of the year; students were not informed of these exams in advance.

Results and Discussion

Twenty-five students who qualified for special education or gifted programs were excluded from our analyses. Furthermore, 39 students who were not present for all initial quizzes, lesson exams, and delayed exams were also excluded to ensure the integrity of our testing schedule. Finally, two additional students who self-reported that they did not have Internet access at home were removed from analyses. Therefore, data from 66 students are reported below, and the general pattern of results remained the same when the 39 absent students were included (see Appendix, Table A-3), confirmed by additional analyses. Means have been collapsed from eight lessons into the five textbook chapters (discussed earlier in the Materials section), and means have been weighted proportionally (as in Experiments 1 and 2).

Initial quiz performance. Initial quiz performance on multiple-choice pretests was 42% with a standard deviation of .10. There was a significant effect of chapter, $F(4, 260) = 21.70$, $\eta_p^2 = .25$, such that pretest performance ranged from 30–52% depending on the chapter. Student performance on the online games was not recorded; however, pretest performance for students who self-reported using the Web site during the course of the experiment

(N = 58) was 42%, whereas pretest performance for students who self-reported never using the Web site during the experiment (N = 8) was 45%, indicating that both groups of students had similar levels of prior knowledge before classroom lectures and self-quizzing via the Internet.

Chapter exam performance: Multiple-choice. Multiple-choice performance as a function of chapter and learning condition is shown in the top two rows of Table 5. Overall, a 2 (tested, nontested) × 5 (chapter) ANOVA revealed that performance was significantly greater for tested items (90%) than nontested items (82%), $F(1, 65) = 29.09$, $\eta_p^2 = .31$, indicating the robust effects of pretesting and/or self-quizzing on retention in a classroom setting. Multiple-choice performance for students who self-reported using the Quia Web site during the experiment was 89% on tested items and 81% on nontested items. In addition, a significant positive correlation between the testing effect (tested performance—nontested performance) for each subject and their self-reported use of the Web site was obtained, $r = .25$, $p = .04$. Thus, testing had a positive effect, and correlational evidence indicates that more frequent quizzing may be associated with higher test performance on unit tests. Multiple-choice performance for the eight students who self-reported never using the Web site during the experiment was 93% on tested items and 88% on nontested items, indicating larger testing effects for students who used the Web site but slightly higher overall performance for students who did not use the Web site. Perhaps these higher performing students felt no need to use the Web site. Chapter exam performance varied across chapters, $F(4, 260) = 161.00$, $\eta_p^2 = .71$, and the benefit of quizzing also varied across chapters, $F(4, 260) = 7.46$, $\eta_p^2 = .10$, such that the benefit was 19% for the ancient Egypt chapter and 3% at a minimum for the ancient India chapter.

Chapter exam performance: Short answer. Short answer performance as a function of chapter and learning condition is shown in the bottom two rows of Table 5. Performance was generally poorer for the short answer test by 6–7% because of the requirement to produce rather than recognize correct answers. A 2 (tested, nontested) × 5 (chapter) ANOVA confirmed that short answer performance for tested items (84%) was significantly greater than nontested performance (75%), $F(1, 65) = 52.44$, $\eta_p^2 = .45$. Short answer performance for students who reported using the Web site during the experiment was 83% on tested items and 73% on nontested items, whereas short answer performance for students who reported never using the Web site during the experiment was 87% on tested items and 83% on nontested items. Similar to multiple-choice chapter performance, these results suggest larger testing effects for students who used the Web site, but slightly higher performance overall for students who did not use the Web site. Again, a significant positive correlation between the testing effect and Web site use was obtained, $r = .39$, $p = .001$. Short answer performance also varied across chapters, $F(4, 260) = 103.19$, $\eta_p^2 = .61$, and there was a significant interaction between chapter and learning condition, $F(4, 260) = 2.86$, $\eta_p^2 = .04$, such that the benefit was 17% for the ancient Egypt chapter and 6% at a minimum for the ancient Greece chapter.

Semester exam performance. Table 6 shows performance on the multiple-choice end-of-the-semester exam for both chapter exam formats (multiple-choice and short answer). Two separate 2 (tested, nontested) × 5 (chapter) ANOVAs were conducted for the end-of-the-semester exam, one for performance following the multiple-choice chapter exam, one for performance following the short answer chapter exam. As shown in the first set of rows of Table 6, multiple-choice performance at the end of the semester following a multiple-choice chapter exam was significantly greater for tested items (74%) than for nontested items (65%), $F(1, 65) = 16.58$, $\eta_p^2 = .20$. Performance also differed across chapters, $F(4, 260) = 31.63$, $\eta_p^2 = .33$, but tested performance was always greater than nontested performance (i.e., there was no significant interaction, $F < 1$). Multiple-choice performance at the end of the semester following a short answer chapter exam (the second pair of rows of Table 6) was similar for tested (70%) and nontested (73%) items, which was not reliable and obviously was opposite the predicted direction of effect, $F = 1.50$. A main effect of chapter was significant, $F(4, 260) = 61.67$, $\eta_p^2 = .49$, but this again just indicates differences due to difficulty of the chapters and/or test items.

In sum, substantial testing effects were obtained on multiple-choice and short answer chapter exams following pretesting and self-quizzing, that is, on exams that occurred relatively soon (from days to weeks) after learning. A significant testing effect for multiple-choice chapter exam items persisted until the end of the semester, for both students who self-reported using the Web site (73% on tested items, 64% on nontested items) and students who did not use the Web site (81% on tested items, 75% on nontested items). The fact that students who used the Web site scored lower than those who did not may suggest that less able students used the Web site in an attempt to improve their performance.

General Discussion

The three experiments reported here revealed positive effects of retrieval practice via quizzing in school classrooms for academic material that students were studying for their course. In addition,

Table 5
Chapter Exam Performance as a Function of Chapter, Test Format, and Learning Condition in Experiment 3

	Ancient Egypt	Ancient Mesopotamia	Ancient India	Ancient China	Ancient Greece	Mean
Chapter Exam: Multiple-Choice						
Tested	.93 (.10)	.92 (.15)	.89 (.18)	.90 (.16)	.85 (.17)	.90 (.10)
Nontested	.78 (.18)	.87 (.18)	.84 (.21)	.84 (.17)	.78 (.19)	.82 (.12)
Chapter Exam: Short Answer						
Tested	.87 (.15)	.81 (.20)	.90 (.19)	.79 (.24)	.83 (.19)	.84 (.13)
Nontested	.74 (.21)	.70 (.26)	.80 (.24)	.72 (.27)	.77 (.19)	.75 (.17)

Note. Overall means have been weighted according to the number of items per chapter. Standard deviations are shown in parentheses.

Table 6
Multiple-Choice Performance on the Semester Exam as a Function of Chapter, Format on the Previous Chapter Exam (MC: Multiple-choice; SA: Short Answer), and Learning Condition in Experiment 3

	Ancient Egypt	Ancient Mesopotamia	Ancient India	Ancient China	Ancient Greece	Mean
End-of-the-Semester (Chapter MC)						
Tested	.70 (.30)	.79 (.28)	.82 (.30)	.70 (.36)	.73 (.31)	.74 (.19)
Nontested	.62 (.37)	.73 (.33)	.59 (.36)	.61 (.35)	.69 (.34)	.65 (.22)
End-of-the-Semester (Chapter SA)						
Tested	.82 (.27)	.70 (.40)	.66 (.32)	.64 (.37)	.66 (.37)	.70 (.21)
Nontested	.85 (.25)	.80 (.32)	.70 (.35)	.70 (.34)	.60 (.37)	.73 (.18)

Note. Overall means have been weighted according to the number of items per chapter. Standard deviations are shown in parentheses.

we obtained positive effects on multiple-choice exams (the ones on which students' grades were based), short answer exams, and a free recall exam somewhat like an essay test. In Experiment 1 we showed that three quizzes with feedback (pretest, posttest, and review test) produced greater performance on chapter exams and on the end-of-the-semester exam relative to items that were not quizzed but were otherwise treated identically in terms of the teacher's lectures and the readings. In Experiment 2, we included a rereading control condition in addition to the tested and nontested conditions to see if reexposure was the sole reason for the benefit of testing. It was not. We replicated the finding that repeated testing improved retention relative to a nontested condition, and we also showed that testing produced greater performance than rereading the material. In fact, performance in the rereading condition did not differ from the nontested condition, in line with other work showing the ineffectiveness of rereading as a study strategy (e.g., Callender & McDaniel, 2009). Experiment 3 extended the positive effects of testing to a short answer test (albeit one that closely followed a multiple-choice test). In Experiment 3 students received only a single pretest, but then they were encouraged to use an Internet-based quizzing Web site outside of class to practice retrieval on their own. Most students availed themselves of this opportunity, and a significant (if modest) correlation showed that students who reported greater use the Web site showed larger gains from testing. Taken together, our results indicated that a test-enhanced learning procedure (McDaniel, Roediger, and McDermott, 2007; Roediger et al., 2010; Roediger & Karpicke, 2006b) can be used to enhance knowledge of material in an actual classroom setting.

At the end of the school year, we gave students questionnaires to survey their attitudes about quizzing via the clicker systems. One worry often expressed by educators is that retrieval practice via quizzing will increase test anxiety and turn students off to school. However, the clicker form of low-stakes quizzing does not appear to have these effects. We found that 97% of students from our study reported that the clicker quizzes increased their learning, 65% reported that the clicker quizzes decreased their test anxiety, and 67% of students reported spending about the same amount of time studying for their social studies class as they did for other classes. Thus, students perceived the use of clickers in class as beneficial and enjoyable. In fact, on days when we did not use the clicker systems, some students complained and asked to use them. In the remainder of the General Discussion, we discuss possible theoretical reasons for our results and describe some limitations of our research.

Possible Theoretical Accounts of Benefits of Quizzing

Why does the retrieval practice (or testing) effect occur for classroom material? Of course, research conducted in the classroom is much better for answering questions of generality of effects (i.e., can laboratory results be extended to the classroom?) than for adjudicating theoretical interpretations (see Roediger & Butler, 2011, for possible mechanisms by which testing benefits learning). Nonetheless, we consider here the two most likely theoretical accounts (which are not themselves inconsistent) and we reject a third account that is often suggested.

To consider the last point first, we can reject the idea that testing benefits performance simply by providing students another opportunity to study material, as has been suggested by Thompson et al. (1978), among others. That hypothesis might work for the results of Experiment 1, which showed that three tests led to better retention than no tests, but the account is ruled out by the results of Experiment 2 that used a rereading comparison condition. Testing produced better retention than did restudying at the same intervals, which shows that the testing effect in the classroom is not simply due to restudying. Roediger and Karpicke (2006a) reviewed considerable laboratory evidence showing that testing provides a greater boost to later retention than does repeated studying, especially when initial testing is followed by feedback and when the final criterial test is delayed. In fact, testing effects often occur even under conditions when production tests are given without feedback and thus, the conditions are tilted toward giving an advantage to the repeated study conditions (e.g., Roediger & Karpicke, 2006b; Wheeler et al., 2003). That is, in repeated study conditions, 100% of the material is, by definition, restudied. However, in repeated test conditions without feedback, students only reexpose themselves to the amount of material that they can produce. Nonetheless, testing often outpaces restudying, especially when the final criterial test is delayed (e.g., Roediger & Karpicke, 2006b).

Two other ideas that have been put forward to account for the testing effect are retrieval effort (and elaboration) and transfer appropriate processing, which are best considered as complementary rather than competitive hypotheses. Gardiner, Craik, and Bleasdale (1973) argued that the effort involved in retrieval might produce the testing effect and provided evidence consistent with this hypothesis (as did Pyc & Rawson, 2009, more recently). In a similar vein, Bjork (1975) and McDaniel and Masson (1985) hypothesized that alternate retrieval routes may be established during testing or that testing may cause elaborative processing in

some other way. Any or all of these mechanisms could be at work in our classroom testing effects. Of course, because we used multiple-choice tests, retrieval effort may seem an unlikely candidate. However, even multiple-choice tests (although not requiring the effort of a production test) do involve considering several response candidates and selecting one over others, processes that are likely to involve more effort than simply rereading the statements as in the control condition in Experiment 2. Still, because retrieval effort is certainly less in multiple-choice tests relative to recall tests, this idea remains tentative in the current context.

A second account for testing effects is that of transfer appropriate processing (e.g., Bransford et al., 1979; Roediger, Gallo, & Geraci, 2002). The basic idea is that students' study activities should match the requirements of the criterial test they will eventually take; more broadly, study processes should ideally instantiate procedures that will be needed when information is used on a later occasion. As discussed in the introduction, students' typical study strategies are to read the text, highlight parts of it deemed important, and then reread the highlighted material when studying for a test. This tactic provides for fluent reprocessing of the material and leads to (unwarranted) confidence that it can be retrieved, but rereading does not permit students to practice retrieving the material, which is of course what will be required on the test and will also be necessary in applying knowledge outside the context of the classroom. Quizzing requires students to practice retrieval, to practice accessing information as well as to reencode it. When students practice in ways required by the later test, they will do better on that test (e.g., Blaxton, 1989; Morris, Bransford, & Franks, 1977).

Either or both theoretical viewpoints outlined above can provide a general account of our results. It may well be that in practicing retrieval via tests, as a means to study and to learn, also requires effort and elaboration of the material. In addition, testing also improves students' metacognitive awareness—testing permits them to discover what they know well enough to retrieve and what material requires further study (Metcalfe & Finn, 2008).

Further research will be necessary to examine these ideas and to unravel the puzzle of why testing enhances learning in the classroom. We have made a start in this direction in two other articles (McDaniel, Agarwal, et al., 2011; McDaniel, Thomas, et al., 2011). For example, in McDaniel, Agarwal, et al. (2011) students in middle school science classes received quizzes according to various schedules involving one quiz (prequiz only, postquiz only, review-quiz only), two quizzes with the various combinations, or all three quizzes. Rather surprisingly, prequizzes did not seem to enhance retention on the later criterial test (either alone or in combination with other quizzes; the review quizzes seemed most critical for retention on the criterial tests given later). Determining the most propitious ways to use quizzing as a learning activity remains a target for future research.

Limitations

Our results are among the first to show that quizzing can be integrated into a classroom context as a normal part of a course over most of a school year to enhance academic performance (see also McDaniel, Agarwal, et al., 2011; McDaniel, Thomas, et al., 2011). The multiple-choice chapter tests we used as our dependent measures were the ones for which students received grades in the course. In our within-student design, we were able to show in all three experiments that students improved from roughly a B− average on nonquizzed items to an A− or A average on quizzed items. If this had been a true educational intervention rather than an experiment, we would have quizzed as much material as possible. Our quizzing effect also extended to exams at the end of the semester. As noted in the introduction, mastering a body of factual knowledge is critical in practically every academic discipline, so uncovering ways to improve the process can only lead to better educational practices. Here we address three possible limitations or criticisms of our results.

First, a critic could complain that our effects were not terribly large, usually around 10% on chapter tests. We have already countered this point briefly when discussing the results of Experiment 1, but the same point can be made for all three experiments. Because baseline (nontested) performance is relatively high on the multiple-choice tests following the lectures and reading (usually about 80% or slightly greater), the range for possible improvement in the quizzing conditions is only about 20%. Despite this constricted range, we showed large relative improvements on chapter tests (the ones that counted toward students' grades) in all three experiments. For example, as noted when discussing Experiment 1, given the baseline of 81%, the most successful possible treatment could only raise performance by 19% and our procedure lifted performance 13/19ths of the way, for 68% improvement. The proportion gain from quizzing was 47% and 44% for Experiments 2 and 3, respectively. Thus, what may look like a rather small gain is an important one in terms of the grading scale used in the classroom. We raised student grades from roughly B− to A−, using the teacher's typical grading scheme. Thus, the effects we produced were really quite healthy given the high baseline performance.

A second issue is that our testing effects were diminished when we reexamined students' knowledge and performance at the end of the semester on material they had learned earlier. We believe it is possible that greater benefits after a long delay may be obtained in future research by using several different means (besides the obvious one of using more items in the final assessment to achieve more stable results). Our initial quizzes were all given the week the teacher covered the material in class; students took a pretest, posttest, and then review test before the criterial chapter test some time later. Thus, the three testing opportunities were massed within one week. If we were able to space quizzes and tests over the entire school year, giving refresher quizzes as it were, then we believe we could maintain the gains from testing, not only at the end of the semester, but also at the end of the school year. However, this statement is a largely promissory note for future research.

A third issue that needs to be addressed is that the multiple-choice quizzes we used for students to practice retrieval involved the same items (albeit with order of alternatives randomized anew) as on the criterial test. A critic could complain that we are merely "teaching to the test" using a "drill and kill" procedure. We make several points in response. First, we showed that quizzing produced better recall of facts on a free recall test in Experiment 1 and on a short answer test in Experiment 3. Second, other research has shown that retrieval practice can also confer gains on transfer tests of information asked in different ways or even to entirely different types of problems with the same formal structure (see Butler, 2010; Rohrer, Taylor, & Sholar, 2010). Third, in another in our

series of middle school experiments, we have shown that quizzing on middle school facts leads to greater transfer on application questions (McDaniel, Thomas, et al., 2011). We should also note that our practice quizzes were limited to use of multiple-choice questions in this series of experiments, but we expect to find greater gains when we use retrieval practice in other forms of testing because laboratory research has shown that production tests followed by feedback produce greater testing effects than do multiple-choice tests with feedback as in the present research (see Kang et al., 2007).

Conclusion

Three experiments showed retrieval practice (testing) effects in middle school classrooms with actual course content, effects that lifted students' performance by a letter grade and that were maintained for several months in some cases. Repeated quizzing led to better performance than did repeated reading. Thus, the current results confirm a wealth of laboratory research showing that the act of taking a test not only measures learning but also changes it, improving performance on later tests.

References

Abbott, E. E. (1909). On the analysis of the factors of recall in the learning process. *Psychological Monographs, 11,* 159–177. doi:10.1037/h0093018

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology, 22,* 861–876. doi:10.1002/acp.1391

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

Blaxton, T. A. (1989). Investigating dissociations among memory measures: Support for a transfer-appropriate processing framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 657–668. doi:10.1037/0278-7393.15.4.657

Bransford, J. D., Franks, J. J., Morris, C. D., & Stein, B. S. (1979). Some general constraints on learning and memory research. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory* (pp. 331–354). Hillsdale, NJ: Erlbaum.

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

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 1118–1133. doi:10.1037/a0019902

Callender, A. A., & McDaniel, M. A. (2009). The limited benefits of rereading educational texts. *Contemporary Educational Psychology, 34,* 30–41. doi:10.1016/j.cedpsych.2008.07.001

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology, 23,* 760–771. doi:10.1002/acp.1507

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

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104,* 268–294. doi:10.1037/0096-3445.104.3.268

Duchastel, P. C., & Nungester, R. J. (1982). Testing effects measured with alternate test forms. *Journal of Education Research, 75,* 309–313.

Gardiner, J. M., Craik, F. I. M., & Bleasdale, F. A. (1973). Retrieval difficulty and subsequent recall. *Memory & Cognition, 1,* 213–216. doi:10.3758/BF03198098

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology, 6*.

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. doi:10.1016/S0022-5371(71)80029-4

Glass, A. L. (2009). The effect of distributed questioning with varied examples on exam performance on inference questions. *Educational Psychology, 29,* 831–848.

Hunt, R. R., & McDaniel, M. A. (1993). The enigma of organization and distinctiveness. *Journal of Memory and Language, 32,* 421–445. doi:10.1006/jmla.1993.1023

Izawa, C. (1971). The test-trial potentiating model. *Journal of Mathematical Psychology, 8,* 200–224. doi:10.1016/0022-2496(71)90012-5

Jones, H. E. (1923−1924). The effects of examination on the performance of learning. *Archives of Psychology, 10,* 1–70.

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19,* 528–558. doi:10.1080/09541440601056620

Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student learning: Do students practice retrieval when they study on their own? *Memory, 17,* 471–479. doi:10.1080/09658210802647009

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science, 319,* 966–968. doi:10.1126/science.1152408

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General, 138,* 469–486. doi:10.1037/a0017341

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* Cambridge, UK: Cambridge University Press.

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General, 138,* 449–468. doi:10.1037/a0017350

Mandler, G. (1967). Organization and memory. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 1, pp. 328–372). New York, NY: Academic Press.

Mayer, R. E. (2008). *Learning and instruction* (2nd ed.). Upper Saddle River, NJ: Merrill Prentice Hall.

Mayer, R. E. (2010). *Applying the science of learning.* Upper Saddle River, NJ: Pearson Merrill Prentice Hall.

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology, 103,* 399–414.

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., & Masson, M. E. J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 370–384.

McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14,* 200–206.

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., Roediger, H. L., & McDermott, K. B. (2011). Quizzing promotes transfer of target principles in middle school science: Benefits on summative exams. Manuscript submitted for publication.

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review, 15,* 174–179.

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

Paivio, A. (1969). Mental Imagery in associative learning and memory. *Psychological Review, 76,* 241–263.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60,* 437–447.

Rawson, K. A., & Kintsch, W. (2005). Rereading effects depend upon time of test. *Journal of Educational Psychology, 97,* 70–80.

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

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Science, 15,* 20–27.

Roediger, H. L., Gallo, D. A., & Geraci, L. (2002). Processing approaches to cognition: The impetus from the levels of processing framework. *Memory, 10,* 319–332.

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

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 233–239.

Sones, A. M., & Stroud, J. B. (1940). Review, with special reference to temporal position. *Journal of Educational Psychology, 31,* 665–676.

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

Swenson, I., & Kulhavy, R. W. (1974). Adjunct questions and the comprehension of prose by children. *Journal of Educational Psychology, 66,* 212–215.

Thompson, C. P., Wenger, S. K., & Bartling, C. A. (1978). How recall facilitates subsequent recall: A reappraisal. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 210–221.

Tulving, E. (1962). Subjective organization in free recall of "unrelated" words. *Psychological Review, 69,* 344–354.

Tulving, E. (1968). Organized retention and cued recall. In H. J. Klausmeier & G. T. O'Hearn (Eds.), *Research and development towards the improvement of education.* (pp. 3–13). Madison, WI: Dembar Educational Research Services.

Ward, D. (2007). eInstruction: Classroom Performance System [Computer Software]. Denton, TX: EInstruction Corporation.

Wheeler, M. A., Ewers, M., & Buonanno, J. F. (2003). Different rates of forgetting following study versus test trials. *Memory, 11,* 571–580.

Wheeler, M. A., & Roediger, H. L. (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science, 3,* 240–245.

Willingham, D. T. (2009). *Why don't students like school?* San Francisco, CA: Jossey-Bass.

Zaromb, F. M., & Roediger, H. L. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & Cognition, 38,* 995–1008.

Appendix

Table A-1
Initial Quiz, Chapter Exam, and Semester Exam Performance as a Function of Learning Condition for All Students Except Gifted and Special Education Students in Experiment 1

	Pre-Test	Post-Test	Review Test	Chapter Exam: Free Recall	Chapter Exam: Multiple-Choice	End-of-the-Semester
Tested	.41 (.09)	.91 (.07)	.91 (.07)	.29 (.11)	.89 (.11)	.76 (.14)
Nontested				.19 (.10)	.78 (.14)	.66 (.17)

Note. Data include all students except gifted and special education students (N = 118, though the number of subjects and items in each cell varies somewhat). Overall means have been weighted according to the number of items per chapter. Standard deviations are shown in parentheses.

Table A-2
Initial Quiz, Chapter Exam, and Semester Exam Performance as a Function of Learning Condition for All Students Except Gifted and Special Education Students in Experiment 2

	Pre-Test	Post-Test	Review Test	Chapter Exam	End-of-the-Semester
Tested	.42 (.12)	.89 (.10)	.86 (.08)	.85 (.22)	.59 (.20)
Read				.78 (.22)	.53 (.21)
Nontested				.77 (.22)	.55 (.18)

Note. Data include all students except gifted and special education students (N = 119, though the number of subjects and items in each cell varies slightly). Overall means have been weighted according to the number of items per chapter. Standard deviations are shown in parentheses.

Table A-3
Initial Quiz, Chapter Exam, and Semester Exam Performance as a Function of Format on the Previous Chapter Exam (MC: Multiple-choice; SA: Short Answer) and Learning Condition for All Students Except Gifted and Special Education Students in Experiment 3

	Pre-Test	Chapter Exam: Multiple-Choice	Chapter Exam: Short Answer	End-of-the-Semester (Chapter MC)	End-of-the-Semester (Chapter SA)
Tested	.41 (.10)	.88 (.10)	.81 (.14)	.74 (.19)	.69 (.21)
Nontested		.81 (.12)	.73 (.17)	.65 (.20)	.71 (.18)

Note. Data include all students except gifted and special education students (N = 103, though the number of subjects and items in each cell varies slightly). Overall means have been weighted according to the number of items per chapter. Standard deviations are shown in parentheses.

Received April 20, 2010
Revision received March 28, 2011
Accepted April 21, 2011 ■