Both Multiple-Choice and Short-Answer Quizzes Enhance Later Exam Performance in Middle and High School Classes

Kathleen B. McDermott, Pooja K. Agarwal, Laura D'Antonio, Henry L. Roediger, III, and Mark A. McDaniel

Washington University in St. Louis

Author Note: This research was supported by Grant R305H060080-06 and Grant R305A110550 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, Ed Settles and Gina Segobiano, Columbia Middle School principal Roger Chamberlain, Columbia High School principals Mark Stuart and Jason Dandurand, teachers Teresa Fehrenz and Neal O'Donnell, all of the 2009-2010 and 2011-2012 7th grade students, 2011-2012 high school students and their parents. We also thank Jessye Brick and Allison Obenhaus for their help preparing materials and testing students, and Jane McConnell, Brittany Butler, Kari Farmer, and Jeff Foster for their assistance throughout the project.

Correspondence:
Kathleen McDermott
Department of Psychology, CB1125
Washington University in St. Louis
One Brookings Drive
St Louis MO 63130-4899
314.935.8743
kathleen.mcdermott@wustl.edu

Abstract

Practicing retrieval of recently-studied information enhances the likelihood of the learner retrieving that information in the future. We examined whether short-answer and multiple-choice classroom quizzing could enhance retention of information on classroom exams taken for a grade. In 7th grade science and high school history classes, students took intermittent quizzes (short-answer or multiple-choice, both with correct-answer feedback) on some information, whereas other information was not initially quizzed but received equivalent coverage in all other classroom activities. On the unit exams and on an end-of-semester exam, students performed better for information that had been quizzed than that not quizzed. An unanticipated and key finding is that the format of the quiz (multiple-choice or short-answer) did not need to match the format of the criterial test (e.g., unit exam) for this benefit to emerge. Further, intermittent quizzing cannot be attributed to intermittent re-exposure to the target facts: A restudy condition produced less enhancement of later test performance than did quizzing with feedback. Frequent classroom quizzing with feedback improves student learning and retention, and multiple choice quizzing is as effective as short answer quizzing for this purpose.

Both Multiple-Choice and Short-Answer Quizzes Enhance Later Exam Performance in Middle and High School Classes

At all levels of education, instructors use classroom quizzes and tests to assess student learning. Laboratory studies demonstrate that tests for recently-learned information are not passive events, however. The assessments themselves can affect later retention. Specifically, attempting to retrieve information can—even in the absence of corrective feedback—enhance the likelihood of later retrieval of that information, relative to a case in which the information is not initially tested (e.g., Carpenter & DeLosh, 2006; Hogan & Kintsch, 1971; McDaniel & Masson, 1985; see McDermott, Arnold, & Nelson, in press, and Roediger & Karpicke, 2006a, for reviews of this phenomenon, known as the testing effect).

Might educators use this knowledge to enhance student learning? That is, could frequent low-stakes testing be used within normal classroom procedures to enhance retention of important classroom material? Laboratory studies are suggestive but insufficient for making recommendations. The typical laboratory study presents a set of information once; this situation differs markedly from the learning done in classrooms, in which integrated content is encountered repeatedly not just within the classroom itself, but also in homework and reading assignments. Further, the typical retention intervals in a class setting are longer than those in laboratory studies. Hence, laboratory experiments are highly suggestive but are insufficient for making definitive recommendations regarding classroom procedures.

Some studies have shown testing effects within classroom settings (Carpenter, Pashler, & Cepeda, 2009; Duchastel & Nungester, 1982; Sones & Stroud, 1940; Swenson & Kulhavy, 1974), although only a few have done so with actual course assessments used for grades in college classrooms (McDaniel, Wildman, & Anderson, 2012) and middle school classrooms (McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011; McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013; Roediger, Agarwal, McDaniel, & McDermott, 2010). These experiments reveal that low-stakes multiple-choice quizzes with immediate correct-answer feedback can indeed enhance student learning for core course content, as revealed in regular, in-class unit exams. For example, Roediger et al. (2011) found in three experiments that students in a 6th grade social studies class were more likely to correctly answer questions on their chapter exams and end-of-semester exams if the information had appeared on in-class multiple-choice quizzes (relative to situations in which the information had not been quizzed or had been restudied). Similarly, in an 8th grade science classroom, McDaniel et al. (2011) showed robust benefits on unit exams for information that had appeared on a multiple-choice quiz relative to nonquizzed information; students answered 92% of the previously-quizzed questions correctly, relative to 79% of the non-quizzed questions. Further, this benefit carried over to end-of-semester and end-of-year exams.

Laboratory work suggests that the format of quizzing (i.e., multiple-choice or short-answer) might influence the effectiveness in enhancing later retention, although cross-format benefits are seen (Butler & Roediger, 2007; Carpenter & DeLosh, 2006; Glover, 1989; Hogan & Kintsch, 1971; Duchastel & Nungester, 1982). For example, Kang et al. (2007) have shown that when feedback is given, short-answer quizzes

covering recently-read excerpts from *Current Directions in Psychological Science* were more effective than multiple-choice quizzes at boosting performance on tests given 3 days later, regardless of whether that final test was in multiple-choice or short-answer format. A similar experiment in a college course found that short-answer quizzes produced more robust benefits on later multiple-choice exams than did multiple-choice quizzes (McDaniel, Anderson, Derbish, & Morrisette, 2007). Similarly, in a simulated classroom setting, Butler and Roediger (2006) showed an Art History lecture to college students. A month later, students returned to the lab and received a short-answer test. Items that had been tested in short-answer format were remembered best (46%), followed by items that had been tested in multiple-choice format or restudied (both 36%). All three conditions exceeded the no-activity condition, for which items were not encountered after the initial lecture. McDaniel, Roediger, and McDermott (2007) reviewed this emerging literature and concluded that "the benefits of testing are greater when the initial test is a recall (production) test rather than a recognition test" (p. 200).

Although this conclusion rests largely on laboratory studies, there are also theoretical reasons to predict this pattern. In the same way that attempting to retrieve information engages active processing that can enhance later memorability, retrieval tests that engender more effortful, generative processes (e.g., short-answer tests) can enhance later memory more than those that are completed with relative ease (e.g., multiple-choice tests). Bjork (1994; see Bjork & Bjork, 2011, for a recent review) has labeled this as the concept of "desirable difficulties" and suggested that retrieval practice is one such desirable difficulty. For example, interleaving instruction on various topics (instead of encountering them all together) helps retention. Similarly, spacing

learning events in time (instead of massing them together) is helpful for long-term retention, although spacing tends to be less effective for the immediate term. In short, the framework of desirable difficulties and the existing laboratory literature both lead to the prediction that short-answer quizzes might facilitate later test performance more than would multiple-choice quizzes.

From an applied perspective, however, using short-answer quizzes to enhance student learning is likely less attractive to middle- and high-school teachers than using multiple-choice quizzes. Short-answer quizzes require more class time to administer and are more challenging to grade. To the extent that multiple-choice quizzes offer benefits similar to those arising from short-answer quizzes, this would be an important practical point and may enhance the likelihood that teachers will attempt to incorporate quizzing into their classrooms.

Accordingly, one purpose of the present study was to investigate the possibility that with an appropriate procedure, multiple-choice quizzes could produce benefits on later exam performance of the magnitude produced by short-answer quizzes. A standard feature of the studies finding advantages for short-answer relative to multiple-choice quizzes is that only a single quiz was given (e.g., Butler & Roediger, 2006; Kang et al., 2007; McDaniel et al., 2007). In recent experiments a different pattern emerged when students were encouraged to take each quiz four times; multiple-choice quizzes enhanced later exam performance as much as did short-answer quizzes (McDaniel et al, 2012). Several features of that study limit the generalizability of the results, however. First, the students took the quizzes online, whenever they wanted (up to an hour before then exam), and were permitted to utilize the textbook and course notes for the quizzes.

To the extent that students consulted their books or notes to complete the quizzes, differences in retrieval difficulty across short-answer and multiple-choice quizzes would have been eliminated (i.e., no retrieval would be required). Thus, the processing advantage linked to short-answer quizzes may have been undercut with the open-book, on-line quizzing protocol (although open-book quizzes can produce benefits, Agarwal et al., 2008). The quizzes in the present experiments were administered during class and were closed-book, so that responding explicitly required retrieval practice.

Another limiting feature of the McDaniel et al. (2012) study is that the course exams were always in multiple-choice format. The robust effects of multiple-choice quizzes may have arisen in part because the exam question format matched the question format for multiple-choice quizzes but not short-answer quizzes. The idea here is that performance on a criterial test may benefit to the extent that the processes required by that test overlap with the processes engaged during acquisition of the information (Morris, Bransford, & Franks, 1977; Roediger, Gallo & Geraci, 2002). Hence, if quizzes enhance learning, quizzes that require cognitive processing similar to the final, criterial test will be the most beneficial. To explore this issue, in this study we also manipulated the unit exam question formats (short-answer or multiple-choice) to determine whether a match in format is needed to achieve the greatest benefits, and in particular to obtain relatively robust testing effects with multiple-choice quizzes.

A final feature of the McDaniel et al. (2012) protocol that may have fostered relatively good performance for the multiple-choice quizzing procedure (relative to the short-answer quizzing procedure) is that the on-line quizzes could be accessed up to an hour before the exam was administered. No data were available on the interval between

the students' last quiz and the exam, but it is possible that students were repeatedly taking the quizzes shortly before the exam. The more challenging retrieval required by short-answer quizzes (if students were not using the text or notes by the fourth quiz) would possibly not produce better exam performance (than multiple-choice quizzes) with short retention intervals (cf. Roediger & Karpicke, 2006b). In the present study, we remedied this limitation by administering both unit exams and end-of-the semester exams and interspersing the initial quizzes over weeks. (How we interspersed the quizzes differed across experiments and is specified for each experiment in the procedure sections.) Thus, the retention interval between quizzing and final testing was on the order of weeks in these experiments, thereby providing a challenging evaluation of the benefits of repeated multiple-choice quizzing relative to those of repeated short-answer quizzing.

Another important issue addressed in the present study concerns the interpretation of test-enhanced effects reported in authentic classroom experiments. In the published experimental studies conducted in pre-secondary educational contexts (McDaniel et al., 2011; McDaniel et al., 2013; Roediger et al., 2011), only one experiment (Roediger et al., 2011, Exp. 2) included a restudy control condition against which to compare the quizzing conditions. The benefit of quizzing relative to restudy was observed on a chapter exam but disappeared by the end-of-semester exam. In all of the other experiments, constraints imposed by implementing an experiment in the classroom prevented a restudy control. Without a restudy control, the interpretation of the quizzing effects is clouded. Specifically, the effects associated with quizzing could reflect factors intertwined with the quizzing, such as repetition of the target material,

spacing of the repetitions, and review of the target material just prior to the unit exams. The present investigation includes experiments with restudy controls so that factors unrelated to the testing-effect *per se* could be ruled out as alternative interpretations of any benefits of quizzing.

As overview, we implemented experiments within the context of a 7th grade science classroom (Experiments 1a, 1b, 2, and 3) and a high school history classroom (Experiment 4), using normal instructional procedures and classroom content. Importantly, quizzes and unit exams contributed toward students' course grades; as such, these studies speak directly to how quizzing can affect subsequent classroom performance. In all cases, students were given correct-answer feedback immediately after each quiz question.

In Experiments 1a and 1b some items (counterbalanced across students) were encountered on three quizzes prior to a unit exam and end-of-semester exam. The quiz type (multiple-choice, short-answer) was manipulated within-student, as was the format of the unit exam (multiple-choice, short-answer). We asked: How do multiple-choice and short-answer quizzes compare in their efficacy in enhancing classroom learning? And does the answer depend upon the format of the criterial unit exam used to assess learning? As will be shown, the two quizzing methods produced equivalent effects and the type of criterial exam (and whether it matched the low stakes quizzes) did not matter.

Experiment 2 also involved 3 quizzes (short-answer format). The key question was how repeated quizzing would compare to repeated restudying of the target facts (i.e., those tested in the quizzes). Would taking quizzes help relative to simply being re-

presented with the important target material (e.g., seeing an answer key to a quiz without actually taking the quiz) an equivalent number of times? As will be seen, quizzing (with feedback) aids learning more than does restudy of the same information in classroom situations.

Experiment 3 addressed whether quizzing benefits would remain when the specific wording of the questions was changed across initial quizzes and between quizzes and the unit exam and when we scaled back to just two quizzes per topic. To anticipate, quizzing helped later performance on the unit exam even when the wording was changed. Experiment 4 extended the findings from middle school to high school and from science to history, demonstrating the generality of the findings.

Experiment 1a

Method

Participants. One hundred forty-one 7th grade students (*M* age = 12.85 years, 80 females) 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. Eleven students declined to include their data in the analyses.

Design and materials. This experiment constituted a 3 (learning condition: multiple-choice quiz, short-answer quiz, not tested) x 2 (unit exam format: multiple-choice, short-answer) within-subjects design. Course materials from two 7th grade science units were used: earth's water and bacteria. Eighteen items from earth's water and 12 items from bacteria (30 items total) were randomly assigned to the six

conditions, five items per condition, with a different random assignment for each of the six classroom sections. Counterbalances were adjusted to ensure that each item was presented in each initial quiz format twice across the six classroom sections. Items appeared in the same format for each of the three quizzes, although items were counterbalanced across students. For multiple-choice questions, the four answer choices were randomly re-ordered for each quiz, unit exam, and delayed exam. An example of multiple-choice and short-answer questions can be seen in the Appendix. Full materials are available from the authors upon request.

Procedure. A research assistant administered three initial quizzes for each unit: a pre-lesson quiz (before the material was taught) a post-lesson quiz (after the material was taught), and a review quiz (a day before the unit exam). Quizzes occurred 6-14 days apart. To avoid potential teaching bias toward specified items, we arranged for the teacher to leave the classroom during pre-lesson quizzes so that classroom coverage of the material occurred before the teacher had any possibility of knowing which items were in which condition for a given class. She was present during post-lesson quizzes and review quizzes, but there were six classes with a different assignment of items to conditions across classes, and the classroom coverage of the material had already occurred. A combination of a clicker response system (Ward, 2007) and paper-and-pencil worksheets were used to administer the initial quizzes.

For multiple-choice questions on initial quizzes, question stems and four answer choices were projected to a screen at the front of the classroom. The research assistant read the question and answer choices aloud, after which students had 30 seconds to click in their answer. After all students responded, a green checkmark appeared next to

the correct answer, and the research assistant read aloud the question stem and correct answer.

For short-answer questions on initial quizzes, question stems were presented on a projection screen at the front of the classroom, and they were read aloud by the research assistant. Students were allotted 75 seconds per question to write their answer on a sheet of paper, and the research assistant instructed students when 30 seconds and 10 seconds remained. When time expired, students were asked to put down their pencils, at which time the research assistant displayed and read aloud the question stem and ideal answer.

Multiple-choice and short-answer items were intermixed on initial quizzes; order of topic mirrored the order in which items were covered in the textbook and classroom lectures.

Paper-and-pencil unit exams were administered by the classroom teacher the day after the review quiz. Students were allotted the full class period (approx. 45 minutes) to answer all experimental questions, as well as additional questions written by the teacher and not used in the experiment. Students received written feedback from the teacher a few days after completing the unit exam. Multiple-choice and short-answer questions were presented in a mixed random order on unit exams, and all classroom sections received the same order.

A delayed exam was administered at the end of the semester (approx. 1-2 months after unit exams) using the same procedural combination of the clicker response system and paper-and-pencil worksheets used during initial quizzes. Each question was presented in the same format (multiple-choice or short-answer) as on the

unit exams. Items were presented in a mixed random order, and all classroom sections received the same order. Due to classroom time constraints, only a limited number of items (24 total, four per condition) from Experiments 1a and 1b could be included on the delayed exam. Thus, in order to maximize power, data for the delayed exam were pooled across Experiments 1a and 1b, and analyses are presented at the end of Experiment 1b.

The experiment (and all those reported here) was implemented without altering the teacher's typical lesson plans or classroom activities (apart from the introduction of the quizzes). Students were exposed to all the typical information through lessons, homework, and worksheets. The only difference is that a subset of that information also received intermittent quizzing.

Scoring. With the assistance of the teacher, the research assistant created a grading rubric for short-answer questions. A response was coded as correct if it included key phrases agreed upon by the research assistant and teacher; a response was coded incorrect if it did not contain the key phrase. Any ambiguities in scoring were discussed and resolved between the research assistant and teacher. An independent research assistant blind to condition also scored each response; inter-rater reliability (Cohen's kappa) was .94.

Results

Preliminary considerations. Twenty-four students who qualified for special education or gifted programs were excluded from the analyses. The students in the special education program were given considerable assistance outside of the classroom (including some practice quizzes closely matched to the criterial test). The gifted

students were on or near ceiling on the quizzes and chapter tests, even in the control condition.

In addition, 61 students who were not present for all quizzes and exams across Experiments 1a and 1b were excluded from our analyses, to enable us to combine data from these two experiments for the delayed semester exam (see Experiment 1b). The pattern of results remained the same with all present and absent students included, however (see Appendix, Table A-1 for data from all present and absent students). Thus, 45 students contributed data to the present analyses. Given our primary interest in the effects of initial quiz and final test question format, analyses have been collapsed over the two science units, and means for each subject were calculated as the number of questions answered correctly out of the total number of questions (30) across the two units of material. All results in this study were significant at an alpha level of .05 unless otherwise noted.

Initial quiz performance. Average performance on the initial quizzes is displayed in Table 1. In general, initial quiz performance increased from the pre-lesson quiz (26%, 95% CI [.23, .29]) to the post-lesson quiz (58%, 95% CI [.53, .63]) and review quiz (75%, 95% CI [.71, .79]). In addition, students answered correctly on the multiple-choice quiz more often than on short-answer quizzes (66% and 40%, respectively; 95% CIs [.62, .69] and [.36, .45], respectively). A 3 (quiz type: pre-lesson quiz, post-lesson quiz, review quiz) x 2 (initial quiz format: multiple-choice, short-answer) repeated measures analysis of variance (ANOVA) confirmed significant main effects of initial quiz format, $F(1, 44) = 120.15$, $p < .001$, $\eta_p^2 = .73$ and quiz type, $F(2, 88) = 338.26$, $p < .001$, $\eta_p^2 = .89$, with no significant interaction, $F(2, 88) = 2.27$, $p =$

.109, η_p^2 = .05. As can be seen in Table 1, students made similar gains in short-answer performance from pre-lesson quiz to post-lesson quiz to review quiz as they did for multiple-choice performance across the three quizzes (on average, about a 25 percentage point gain between successive quizzes for both initial quiz formats).

Unit exam performance. Average unit exam performance is displayed in Figure 1. In general, students performed best on the unit exam for questions that had occurred on multiple-choice quizzes (79%; 95% CI [.74, .83]), next best for items that had appeared on the short-answer quizzes (70%, 95% CI [.65, .76]), and worst on items not previously tested (64%, 95% CI [.59, .69]), demonstrating the large benefits of quizzing on end-of-the-unit retention. A 3 (learning condition: multiple-choice quiz, short-answer quiz, not tested) x 2 (unit exam format: multiple-choice, short-answer) repeated measures ANOVA revealed significant main effects of learning condition, $F(2, 88) = 12.32$, $p < .001$, $\eta_p^2 = .22$, and unit exam format, $F(1, 44) = 22.21$, $p < .001$, $\eta_p^2 = .34$, qualified by a significant interaction, $F(2, 88) = 5.25$, $p = .007$, $\eta_p^2 = .11$. We now examine the locus of the interaction by considering each of the unit exam formats in turn. To preview, performance on multiple-choice exam questions was not reliably affected by quizzing, whereas performance on short-answer exam questions was robustly enhanced by the initial quizzes.

***Multiple-choice unit exam*.** A one-way ANOVA on final multiple-choice performance (learning condition: multiple-choice quiz, short-answer quiz, not tested) revealed no significant effect of learning condition, $F(2, 88) = 1.53$, $p = .223$, $\eta_p^2 = .034$. That is, although performance on the multiple-choice unit exam was numerically greater when initial quizzes had been multiple-choice (81%, 95% CI [.76, .87]) than when initial

quizzes had been short-answer or when the items had not been quizzed (both 75%, both 95% CIs [.69, .81]), this difference among means was not reliable.

***Short-answer unit exam*.** A one-way ANOVA on final short-answer performance (learning condition: multiple-choice, short-answer, not tested) revealed a significant effect of learning condition, $F(2, 88) = 16.88$, $p < .001$, $\eta_p^2 = .27$. Initial multiple-choice quizzes and initial short-answer quizzes produced greater short-answer exam performance (76% and 65%, 95% CIs [.70, .82] and [.57, .73], respectively) than seen on not tested items (52%, 95% CI [.45, .60]), $t(44) = 5.91$, $p < .001$, $d = 1.06$, 95% CI [.62, 1.50] and $t(44) = 3.23$, $p = .002$, $d = .50$, 95% CI [.08, .92], respectively. In addition, initial multiple-choice quizzes produced significantly greater short-answer exam performance than initial short-answer quizzes, $t(44) = 2.54$, $p = .015$, $d = .47$, 95% CI [.05, .89].

Discussion

In summary, student performance increased across the quizzes (pre-lesson; post-lesson; review), demonstrating that they progressively learned the material. The key question, though, was: Did the initial quizzes enhance performance on the later unit exam?

When the unit exam was in short-answer format, the answer is clear: taking quizzes (with feedback) enhanced later performance. This was especially true when the quizzes had been in multiple-choice format (perhaps due to higher levels of quiz performance), but the benefit appeared for both multiple-choice and short-answer quizzes. When the unit exam was in multiple-choice format, no significant differences occurred among the three learning conditions (multiple-choice quizzes, short-answer

quizzes, not previously tested), although the multiple-choice quizzing condition produced numerically greater performance.

Experiment 1a demonstrated that a match in question format is not necessary for students to benefit from in-class quizzing. That is, the quiz question does not have to be in the identical format as is used on the unit exam. Indeed, the items that showed the biggest advantage from the quizzes were the items tested initially in a multiple-choice format and later tested with short-answer questions. These findings extend prior work by demonstrating that repeated closed-book multiple-choice quizzes taken intermittently in the days and weeks prior to classroom exams enhance performance on the later multiple-choice and short-answer unit exams.

Experiment 1b was designed to replicate and extend these basic findings of the power of quizzing. In other work conducted in parallel with Experiment 1a, we have shown that pre-lesson tests are ineffective at enhancing student learning in the classroom (McDaniel et al., 2011). In order to maximize learning within classroom time constraints, we re-ordered the placement of the three quizzes. Instead of the first quiz occurring before the teacher lectured on the topic, we placed the initial quiz after the lesson. Hence, students received two post-lesson quizzes and a review quiz prior to the unit exam. Again, we examined how multiple-choice and short-answer quizzes (with feedback) would affect long-term retention of classroom material and whether the answer depends upon the format of the criterial test.

Experiment 1b

Method

Participants. The same 141 students who participated in Experiment 1a also participated in Experiment 1b, which occurred later in the fall semester of the same academic year

Design and materials. The same design from Experiment 1a was used for Experiment 1b. Course materials from three 7th grade science units were used: protists and fungi, plant reproduction and processes, and cells. Twelve items from protists and fungi, 18 items from plant reproduction and processes, and 24 items from cells (54 items total) were randomly assigned to the six conditions, nine items per condition, with a different random assignment for each of the six classroom sections.

Procedure. Procedures were similar to those of Experiment 1a, except for the removal of the pre-lesson quiz. After being taught the material students received two post-lesson quizzes and a review quiz. The first post-lesson quiz occurred 1-3 days after introduction of lesson material, and post lesson and review quizzes occurred 1-6 days apart. The review quiz always occurred the day before the unit exam. All other procedures from Experiment 1a were followed for Experiment 1b.

Scoring. Scoring procedures remained the same as for Experiment 1a. Inter-rater reliability (Cohen's kappa) for short-answer responses was .93.

Results

As discussed previously, the same students excluded from analysis in Experiment 1a were excluded from analysis in Experiment 1b, which allowed us to aggregate data from these students for the delayed semester exam analysis. Even so, the general pattern of results remained the same with all present and absent students included (see Appendix, Table A-2 for data from all present and absent students). Thus,

the remaining analyses include data from the same 45 students as in Experiment 1a. Given our primary interest in the effects of initial quiz and final test question format, analyses have been collapsed over the three science units, and means for each subject were calculated as the number of items correct out of the total number of items (54 items) across the three units of material.

Initial quiz performance. Average initial quiz performance is displayed in Table 2. In general, initial quiz performance increased from the first post-lesson quiz (46%, 95% CI [.42, .49]) to the second post-lesson quiz (58%, 95% CI [.55, .62]) and review quiz (71%, 95% CI [.67, .74]). In addition, students tended to answer correctly more often on the multiple-choice quizzes (78%, 95% CI [.75, .81]) than short-answer quizzes (38%, 95% CI [.34, .43]). A 3 (quiz type: post-lesson quiz 1, post-lesson quiz 2, review quiz) x 2 (initial quiz format: multiple-choice, short-answer) repeated measures ANOVA confirmed significant main effects of quiz type, $F(2, 88) = 241.63$, $p < .001$, $\eta_p^2 = .85$, and initial quiz format, $F(1, 44) = 325.15$, $p < .001$, $\eta_p^2 = .88$, qualified by a significant interaction, $F(2, 88) = 14.61$, $p < .001$, $\eta_p^2 = .25$. As can be seen in Table 2, students made greater gains in short-answer performance from post-lesson quiz 1 to post-lesson quiz 2 to the review quiz (approximately 16 percentage point gain between quizzes) than in multiple-choice performance (approximately 9 percentage point gain from quiz to quiz). This pattern is likely attributable to the fact that multiple-choice items were answered quite well on the first post-lesson quiz (69%, 95% CI [.64, .73]), so there was less room on the scale for these items to demonstrate improvement.

Unit exam performance. Average unit exam performance is displayed in Figure 2. Overall, unit exam performance was greater following initial multiple-choice (72%,

95% CI [.68, .77]) and short-answer quizzes (73%, 95% CI [.69, .77]), compared to not tested items (55%, 95% CI [.50, .60]), demonstrating the large benefits of quizzing on end-of-the-unit retention. A 3 (learning condition: multiple-choice quiz, short-answer quiz, not tested) x 2 (unit exam format: multiple-choice, short-answer) repeated measures ANOVA revealed a significant main effect of learning condition, $F(2, 88) = 45.14$, $p < .001$, $\eta_p^2 = .51$. In addition, there was a significant main effect of unit exam format, $F(1, 44) = 191.98$, $p < .001$, $\eta_p^2 = .81$, confirming that students answered more multiple-choice items correctly (80%, 95% CI [.76, .83]) than short-answer items (54%, 95% CI [.49, .59]). There was no significant interaction of learning condition and unit exam format, $F(2, 88) = .65$, $p = .523$, $\eta_p^2 = .02$. Initial quizzes enhanced unit exam performance (i.e. a testing effect was observed). A match between quiz format and unit exam format was not necessary for this benefit, nor did the match enhance the benefit obtained from the initial quizzes.

End-of-semester exam performance. As described earlier, due to a limited number of items, data for a delayed exam administered at the end of the semester (1-2 months after unit exams) were pooled across Experiments 1a and 1b and are displayed in Figure 3. Not surprisingly, the likelihood of getting multiple-choice items correct (60%, 95% CI [.55, .65]) was greater than that for short-answer (37%, 95% CI [.30, .43]). Further, performance on the delayed exam was greater following multiple-choice (57%, 95% CI [.50, .64]) and short-answer quizzes (51%, 95% CI [.44, .57]) than for items that had not been initially quizzed (but that had been tested once on the unit exam, 38%, 95% CI [.31, .44]). That is, a testing effect was observed after a long delay. A 3 (learning condition: multiple-choice quiz, short-answer quiz, not tested) x 2 (unit exam

format: multiple-choice, short-answer) repeated measures ANOVA confirmed a significant main effect of learning condition, $F(2, 88) = 16.25$, $p < .001$, $\eta_p^2 = .27$, a significant main effect of unit exam format, $F(1, 44) = 78.58$, $p < .001$, $\eta_p^2 = .64$, and a significant interaction, $F(2, 88), = 3.74$, $p = .028$, $\eta_p^2 = .08$.

Simple main effects tests showed a significant effect of learning condition on end-of-semester exams for both multiple-choice, $F(2, 88) = 3.35$, $p = .04$, $\eta_p^2 = .071$, and short-answer final exam questions, $F(2,88) = 16.68$, $p < .001$, $\eta_p^2 = .275$. For the multiple-choice final exam questions, students performed better for items that had been quizzed in the short-answer format than those not quizzed, $t(44) = 2.25$, $p = .029$, $d = .44$, 95% CI [.02, .86]. The 10 percentage point benefit for items quizzed in the multiple-choice format fell short of statistical significance, $t(44) = 1.98$, $p = .054$, $d = .39$, 95% CI [.00, .81], as did the difference between quiz types (multiple-choice or short-answer), $t(44) = .363$, $p = .718$, $d = .07$, 95% CI [.00, .48]. For the questions assigned to short-answer on the final exam, students did best when the initial quizzes had been in multiple-choice format compared to short answer [$t(44) = 2.51$, $p = .016$, $d = .44$, 95% CI [.02, .86]] or not quizzed [$t(44) = 5.81$, $p < .001$, $d = 1.06$, 95% CI [.62, 1.50]]. Students did next best when the quizzes had been in short-answer format and least well for items not previously quizzed, $t(44) = 3.44$, $p = .001$, $d = .50$, 95% CI [.08, .92].

Discussion

To review, in Experiment 1b quiz performance increased from the first post-lesson quiz to the second post-lesson quiz to the review quiz. The key question was how these quizzes would affect retention on the later unit exam and the end-of-semester exam.

On the unit exam, students performed better with information that had appeared on the quizzes than information that had not been quizzed (16 percentage point gain and 21 percentage point gain for unit exam items tested with multiple-choice and short-answer, respectively). Students benefitted greatly from the quizzes.

Further, the exact format of the quizzes did not matter. Students benefitted as much when the quiz and unit exam formats mismatched as when they matched. What mattered was that quizzing with correct answer feedback had occurred. A similar pattern was seen on the end-of-semester exam, although here there was evidence that initial multiple-choice quizzes were especially beneficial when the end of semester question was short-answer. A peculiar result occurred in Experiment 1a where we found no testing effect on the multiple-choice unit test. Because many prior experiments have obtained such an effect on multiple-choice tests (e.g., McDaniel et al., 2011, Roediger et al., 2012) and because we obtained the finding in Experiment 1b, we suggest that the lack of effect in Experiment 1a was likely a Type II error (and note that the numerical difference in 1a was in the direction of showing a testing effect).

These data are especially interesting in light of the performance gap between learning conditions on quiz 1 (Table 1). That is, students do much better on their first multiple choice quiz than their first short answer quiz, a finding that accords with typical classroom findings (and will be seen across all our experiments). Despite this gap, the provision of feedback makes the quizzes equally effective in boosting later memory for the quizzed information. This outcome may point to a role for test-enhanced learning in the classroom; that is, students learn better from presentations when they are preceded by a test than when they are not (Arnold & McDermott, 2013a, 2013b), possibly

because subsequent encounters with the information remind them of their prior test experience, with this recursive reminding enhancing subsequent memory (Nelson et al., 2013).

One potential concern with these findings is that it is not the quizzing *per se* but the selective and spaced re-exposure to the information that aids later performance. Prior work in the laboratory suggests this concern would not account for the present findings (Butler 2010; Roediger & Karpicke, 2006b); restudying information is generally not as beneficial as attempting to retrieve target information from memory. Most pertinent to the present results, Roediger et al. (2011, Exp. 2) have shown in a 6th grade social studies class that multiple-choice quizzing course content led to better performance on multiple-choice chapter exams (*M* = .91) than did re-studying the material, which did not differ from the non-tested control condition (*M* = .83 and .81, respectively).

Also relevant is a study by Carpenter et al. (2009), who drew target facts from an 8th grade US History class and had students take a short-answer quiz with feedback (15 questions), restudy the facts (15 of them), or neither (15 facts). Nine months later, the full set of facts was tested. Students performed quite poorly but did slightly better with facts previously quizzed (10% correct using lenient scoring) than those restudied (7% correct) or those not reviewed (5%). This situation differs from the present one in several key ways; most important is that the final test was not part of the regular classroom activities, was unexpected, and did not count toward the student's grade.

These two studies are the only two classroom experiments that have incorporated a restudy control condition (see too McDaniel et al., 2011; McDaniel et al.,

2013; Roediger et al., 2011, Exps. 1 and 3). Therefore, to confidently conclude that quizzing *per se* is beneficial to learning (exam performances) it is essential to establish that these prior findings are replicable (i.e., that quizzing offers benefits over restudying) and generalize to other classroom contexts.

In Experiment 2, we asked a question similar to that addressed by Roediger et al (2011) concerning the effects of quizzing versus restudying, but here we used short-answer quizzes, short-answer unit and semester-end exams, and the course content was 7th grade science. We examined whether short-answer quizzing would surpass restudying in enhancing later performance on tests. As in Experiments 1a and 1b, some key items from the lesson were quizzed and others were withheld from the quizzes but still taught in the classroom, with the teacher not knowing which items were assigned to which condition for any given class. The new twist here involves a restudying condition. Instead of attempting to answer a short-answer question and then being given feedback, in the control condition students restudied the identical information in statement form. This condition is equivalent to studying the answers to an upcoming test before taking the test.

Experiment 2

Method

Participants. The same students who participated in Experiments 1a and 1b took part in Experiment 2, which was administered in the Spring of the same school year.

Design and materials. Three learning conditions were used in this experiment (quizzed, restudied, not tested), following a within-subjects design. Course materials

from five 7th grade science units were used: motion and momentum; forces and fluids; work, machines, and energy; animals, mollusks, worms, arthropods, and echinoderms; and birds, mammals, and animal behavior. Eighteen items from motion and momentum, 12 items from forces and fluids, 30 items from work, machines, and energy, 30 items from animals, mollusks, worms, arthropods, and echinoderms, and 30 items from birds, mammals, and animal behavior (for a total of 120 items total) were randomly assigned to the three conditions, forty items per condition, with a different random assignment for each of the six classroom sections. Counterbalances were adjusted to ensure that each item was presented in each condition twice across the six classroom sections. All quizzes and exams were completed in a short-answer format (i.e., multiple-choice questions were not used in this experiment). An example of a restudied item can be seen in the Appendix.

Procedure. Similar to Experiment 1b, students received three initial quizzes for each unit (i.e., two post-lesson quizzes and one review quiz), using the same procedural combination of the clicker response system software to display short answer questions and paper-and-pencil worksheets. The first post-lesson quiz was administered 1-3 days after the introduction of lesson material; the post-lesson and review quizzes occurred 1-6 days apart. For the restudy condition, students saw a complete statement (question stem and ideal answer) on the projection screen. Students were asked to follow along as the statement was read aloud by the research assistant. Quizzed and restudied items were presented in a mixed fashion on initial quizzes, in the order in which items were covered in the textbook readings and

classroom lectures. All other procedures from Experiments 1a and 1b were followed for Experiment 2.

Scoring. Scoring procedures remained the same as for Experiments 1a and 1b. Inter-rater reliability (Cohen's kappa) for short-answer responses was .94.

Results

As in Experiment 1a, the twenty-four students who qualified for special education or gifted programs were excluded from analysis. In addition, 47 students who were not present for all quizzes and exams were also excluded from our analyses, but the general pattern of results remained the same with all present and absent students included (see Appendix, Table A-3 for data from all present and absent students). Thus, the remaining analyses include data from 59 students. Given our primary interest in the effects of initial quiz and final test question format, analyses have been collapsed over the five science units, and means for each subject were calculated as the number of items correct out of the total number of items (120 items) across the five units of material.

Initial quiz performance. Average initial quiz performance is displayed in Table 3. Initial quiz performance increased from the first post-lesson quiz (42%, 95% CI [.38, .47]) to the second post-lesson quiz (59%, 95% CI [.54, .64]) and review quiz (74%, 95% CI [.69, .78]), as confirmed by a reliable main effect, $F(2, 116) = 301.84$, $p < .001$, $\eta_p^2 = .84$.

Unit exam performance. Average unit exam performance is displayed in Figure 4. Unit exam performance was greatest for the items that had previously been quizzed (81%, 95% CI [.77, .85]), followed by the restudy (62%, 95% CI [.57, .66]) and not

previously tested (55%, 95% CI [.50, .59]) conditions. A one-way ANOVA confirmed a main effect of learning condition, $F(2, 116) = 154.08$, $p < .001$, $\eta_p^2 = .73$. Planned comparisons confirmed a significant testing effect, such that exam performance for quizzed items was significantly greater than for not tested items, $t(58) = 16.82$, $p < .001$, $d = 1.61$, 95% CI [1.13, 2.08]. Performance on quizzed items was also greater than for restudied items, $t(58) = 12.17$, $p < .001$, $d = 1.15$, 95% CI [.70, 1.59], and performance for restudied items was greater than for items not quizzed, $t(58) = 4.61$, $p < .001$, $d = .39$, 95% CI [.00, .81].

End-of-semester exam performance. Average delayed exam performance is displayed in Figure 5. Again, delayed performance was greatest for the quizzed condition (66%, 95% CI [.62, .71]), followed by performance in the restudy (50%, 95% CI [.44, .56]) and not tested conditions (items not quizzed but that had been tested once on the unit exam, 43%, 95% CI [.38, .49]). A one-way ANOVA confirmed a main effect of learning condition, $F(2, 116) = 35.77$, $p < .001$, $\eta_p^2 = .38$. Planned comparisons confirmed a significant testing effect, $t(58) = 8.07$, $p < .001$, $d = 1.16$, 95% CI [.71, 1.60], a significant effect of quizzing greater than restudying, $t(58) = 6.01$, $p < .001$, $d = .81$, 95% CI [.38, 1.24], and a significant effect of restudying compared to not tested, $t(58) = 2.24$, $p = .029$, $d = .28$, 95% CI [.00, .69].

Discussion

When some key information was quizzed in the classroom and other key information was selectively re-exposed, the quizzed information was retained better. Specifically, relative to restudying the target facts, short-answer quizzing enhanced performance on the unit exam by 19 percentage points and the end-of-semester exam

by 16 percentage points. Both unit exam and end-of-semester exam were short-answer. Selective restudying did aid retention (relative to no exposure), but this effect was much smaller than the quizzing effect. Importantly, our data show that benefits achieved from the repeated quizzing procedure cannot be attributed to simple restudying or to spacing of the re-exposed content (cf. McDaniel et al., 2013).

Experiment 3

In Experiments 1a, 1b, and 2, the identical wording was used in the three quizzes and the unit exam. If the wording is changed across the tests, but the same basic concepts are targeted, will the beneficial effects of quizzing remain? Or are the quizzing effects limited to situations in which the wording is identical from one quiz to the next, to the criterial test? Of course, this issue has important practical considerations, as teachers may be reluctant to repeat questions word-for-word across tests. More importantly, if students are simply memorizing answers to specific questions and not truly learning the material when faced with retrieval practice, its benefits would be very limited. Several laboratory studies have shown this is not so (e.g., Butler, 2010; Carpenter, 2012), but the data from classroom studies is more sparse (see McDaniel et al., 2007, for an experiment with university students, and McDaniel et al., 2013, for experiments with middle school students).

Experiment 3 addressed this issue by changing the question stems between the quizzes and the criterial test so that various wordings were used in each case. For example, for the concept of *spontaneous generation* one question was “Hundreds of years ago, people believed life could appear suddenly from nonliving things. What was their mistaken idea known as?” The same concept was targeted on a different quiz as

"What is the idea that living things arise from nonliving sources, such as flies arising from decaying meat?" And on the unit test it was "When frogs appeared in mud puddles after heavy rains, people concluded frogs could sprout from mud in ponds. What is the term for their mistaken belief that life could come from nonliving sources?" Admittedly, this changing of wording represents near transfer and does not test for deeper understanding of the concept. Nonetheless, it does address the concern that students are simply memorizing answers to question stems without a basic understanding of the concepts.

A second goal of Experiment 3 was to ask whether benefits of quizzing, and especially the robust effects of multiple-choice quizzing, would be manifested after only two initial quizzes (instead of the three used previously). A third goal was to reinforce the Experiment 2 finding that repeated exposure to the target content does not explain the quizzing effects reported here by employing a rereading control condition. Finally, the equipment used in this experiment permitted clicker response systems to be used for both short-answer and multiple-choice quizzes, thus equating the quizzes with respect to response modality.

Method

Participants. One hundred fifty-two 7th grade students (*M* age = 12.18 years, 70 females) from the same public middle school as the previous experiments were invited to participate in this study. This experiment took place in a different school year than the prior experiments and therefore involved different students. Assent was obtained from students in accordance with guidelines of the Human Research Protection Office. Twenty-five students declined to include their data in the analyses.

Design and materials. This experiment followed a 4 (learning condition: multiple-choice quiz, short-answer quiz, restudy, not tested) x 2 (unit exam format: multiple-choice, short-answer) within-subjects design. Course materials from five 7th grade science units were used: bacteria; protists and fungi; plant reproduction and processes; animals, mollusks, worms, arthropods, and echinoderms; and birds, mammals, and animal behavior. Sixteen items from the bacteria unit, 16 items from protists and fungi, 24 items from plant reproduction and processes, 24 items from animals, mollusks, worms, arthropods, and echinoderms, and 24 items from birds, mammals, and animal behavior (104 items total) were randomly assigned to the eight conditions, 13 items per condition, with a different random assignment for each of six classroom sections. Counterbalances were adjusted to ensure that each item was presented in each initial quiz format at least once across the six classroom sections. Items appeared in the same format (short-answer or multiple-choice) for each of the quizzes. However, unlike previous experiments, the questions were rephrased for each quiz or exam such that the students were never tested on the same question verbatim more than once. See the Appendix for sample test questions. Full materials are available from the authors upon request.

Procedure. A research assistant administered two initial quizzes for each unit: a post-lesson quiz (after the material was taught), and a review quiz (prior to the unit exam). The post and review quizzes occurred 1 – 12 days apart (*M* = 4.2 days). A procedure very similar to that of Experiments 1a, 1b, and 2 was followed to administer the quizzes with a clicker response system (Ward, 2007). The main difference here was that the clickers had both multiple-choice and short-answer capabilities, eliminating the

need for the paper-and-pencil worksheets for short-answer responses. For multiple-choice questions, students simply selected the letter A – D corresponding to their answer choice. For short-answer questions, students typed their responses using a touch-tone telephone style keypad (i.e., letters A, B, C were located on the same key as the number 2). Students were familiar with the system and few difficulties were encountered.

For multiple-choice items on initial quizzes, students had a maximum of 30 seconds to click in their answers. After all the students responded, a green checkmark was displayed next to the correct answer, and the research assistant read the complete question stem with the answer inserted. For short-answer items on initial quizzes, students were allowed up to 90 seconds to key in their responses. Students were given a 30 second warning and told to finish up their answers. All answers were fewer than 140 characters, and 90 seconds was more than enough time to key in the responses. Most student submitted responses within the first 60 seconds, but occasionally students who were changing or editing their answers utilized the full 90 seconds. Again, the correct answer feedback was displayed and read immediately after each question. For the restudy condition, the same procedure was used as in Experiment 2. All multiple-choice quiz, short-answer quiz, and restudied items were intermixed; order of topic mirrored the order in which items were covered in the textbook.

Paper-and-pencil unit exams were administered by the classroom teacher approximately 2 - 3 days after the review quiz (M = 2.8 days). Students were allotted the full class period (approximately 45 minutes) to answer all experimental questions, as well as additional questions written by the teacher and not used in the experiment.

Students received written feedback from the teacher a few days after completing the unit exam. Multiple-choice and short-answer questions were presented in a mixed random order on unit exams. All classroom sections received the same exam questions, but sometimes they appeared in a different random order to prevent cheating.

Scoring. Short-answer questions required very brief answers, so a simple correct or incorrect coding system was used. For some questions, the teacher decided to give half credit for the purposes of calculating students' exam grades. However, those items were marked incorrect for the coding in our experiment. An independent research assistant blind to condition also scored a random 10% of responses. Inter-rater reliability (Cohen's kappa) was .87.

Results

Preliminary considerations. Eleven students who qualified for special education programs were excluded from the analyses. In addition, 56 students who were not present for all quizzes and exams for the duration of this experiment were excluded from our analyses. Still, the general pattern of results remained the same with all present and absent students included (see Appendix, Table A-4 for data from all present and absent students). Sixty students were included in the present analyses.

Initial quiz performance. Average performance on initial quizzes is displayed in Table 4. In general, initial quiz performance increased from the post-lesson quiz (56%, 95% CI [.53, .59]) to the review quiz (70%, 95% CI [.66, .73]). In addition, students answered multiple-choice questions correctly more often than short-answer questions (78% and 48%, 95% CIs [.75, .80] and [.45, .52], respectively). A 2 (Quiz Type: post-lesson quiz, review quiz) x 2 (initial quiz format: multiple-choice, short-answer) repeated

measures ANOVA confirmed significant main effects of initial quiz format, $F(1, 59) = 400.31$, $p < .001$, $\eta_p^2 = .87$, and quiz type, $F(1, 59) = 221.41$, $p < .001$, $\eta_p^2 = .79$, qualified by a significant interaction $F(1, 59) = 11.18$, $p = .001$, $\eta_p^2 = .16$. As can be seen in Table 4, students had greater gains in short-answer performance from post-lesson quiz to review quiz (an approximately 16 percentage point gain) than in multiple-choice performance (approximately 11 percentage point improvement from post-lesson to review quiz).

Unit exam performance. Average unit exam performance is displayed in Figure 6. In general, students performed best on the unit exams for questions that had occurred on short-answer quizzes (84%, 95% CI [.81, .87]), next best for items that had appeared on the multiple-choice quizzes (83%, 95% CI [.80, .86]), next best for items restudied (76%, 95% CI [.71, .79]), and worst for items neither restudied nor previously tested (72%, 95% CI [.68, .76]), demonstrating the large benefits of quizzing on end-of-unit retention. A 4 (learning condition: multiple-choice quiz, short-answer quiz, restudy, not tested) x 2 (unit exam format: multiple-choice, short-answer) repeated measures ANOVA revealed significant main effects of learning condition, $F(3, 177) = 45.43$, $p < .001$, $\eta_p^2 = .44$, and unit exam format, $F(1, 59) = 73.91$, $p < .001$, $\eta_p^2 = .56$, qualified by a significant interaction, $F(3, 177) = 4.49$, $p = .005$, $\eta_p^2 = .07$. The effect of learning condition varied slightly across unit exams. We will consider each unit exam format in turn.

Multiple-choice unit exam. A one-way ANOVA on multiple-choice exam performance (learning condition: multiple-choice, short-answer, restudy, not tested) revealed a significant effect of learning condition, $F(3, 177) = 21.19$, $p < .001$, $\eta_p^2 = .26$.

Initial multiple-choice quizzes and initial short-answer quizzes produced greater multiple-choice exam performance (88% and 87%, 95% CIs [.86, .90] and [.85, .90], respectively) than seen on both initially restudied and not tested items (83% and 75%, 95% CIs [.80, .86] and [.71, .79], respectively), *t*s > 2.97, *p*s < .004, *d*s > .48. In addition, performance on items restudied (83%) was significantly greater than the not tested items (75%), $t(59) = 3.82$, $p < .001$, $d = .51$, 95% CI [.09, .93]. The format of initial quiz (multiple-choice or short-answer) did not significantly affect performance, $t(59) = .55$, $p = .59$, $d = .08$, 95% CI [.00, .49].

***Short-answer unit exam*.** A one-way ANOVA on short-answer exam performance (learning condition: multiple-choice, short-answer, restudy, not tested) revealed a significant effect of learning condition, $F(3, 177) = 26.52$, $p < .001$, $\eta_p^2 = .31$. Initial multiple-choice and short-answer quizzes produced greater short-answer exam performance (78% and 81%, 95% CIs [.74, .82] and [.78, .85], respectively) than seen on both restudied and not tested items (68% and 68%, 95% CIs [.62, .73] and [.63, .73]), *t*s > 5.33, *p*s < .001, *d*s > .56. In addition, initial short-answer quizzes produced a slight but significant benefit over initial multiple-choice quizzes (81% and 78%, respectively), $t(59) = 2.01$, $p = .049$, $d = .20$, 95% CI [.00, .61]. Performance on items restudied (68%) was not different from the not tested items (68%), $t(59) = 0.11$, $p = .914$, $d = .01$, 95% CI [.00, .42].

Discussion

The primary findings of Experiment 3 were that changing the wording from quiz to quiz to final test did not remove the testing effect, nor did reducing the number of initial quizzes to two. The effects of word change are important and demonstrate that

students are not merely memorizing answers to the specific wording of questions but in fact are learning the concepts through quizzing.

Students benefitted from both short-answer and multiple-choice quizzes, regardless of the format of the final test. Further, the restudying condition did not produce the same performance benefits seen through quizzing. When the final test was multiple-choice, restudying did produce some benefit, perhaps through increased familiarity with the term. When the final test was short-answer, however, restudying had no benefit relative to the no-test condition.

Experiment 4

Thus far, the studies presented have involved middle school students in science classes. Experiment 4 extends the previous findings into a high school history classroom with the goal of establishing their generalizability to an older sample of students and a different subject matter. As in Experiment 3, the wording of the questions was varied across the quizzes and test in Experiment 4. We did not include a restudy control condition because we had fewer students with whom to work and because Experiments 2 and 3 (and many prior laboratory experiments) have shown that quizzing produces gains greater than that produced by a restudy control condition.

Method

Participants. Seventy-eight 11th and 12th grade high school students (*M* age = 16.21 years, 44 females) from a public high school located in a Midwestern suburban, middle-class community were invited to participate in this study. Assent was obtained from each student in accordance with guidelines of the Human Research Protection Office. Five students declined to include their data in the analyses.

Design and materials. This experiment followed a 3 (learning condition: multiple-choice, short-answer, not tested) x 2 (unit exam format: multiple-choice, short-answer) within-subjects design. Course materials from two American History units were used: 24 items from the Civil Rights unit and 30 items from World War II unit (54 items total) were randomly assigned to each of the six conditions, 9 items per condition, with a different random assignment for each of three classroom sections. Counterbalances were adjusted to ensure that each item was presented in each initial quiz format once across the three classroom sections. Items appeared in the same format for each of the two initial quizzes. However, questions were reworded for each quiz or exam, as was done in Experiment 3. See the Appendix for examples of multiple-choice and short-answer questions. Full materials are available from the authors upon request.

Procedure. A research assistant administered two initial quizzes for each unit: a post-lesson quiz (after the material was taught), and a review quiz (a day before the unit exam). Due to the amount of course material that needed to be covered in a given period of time, the schedule of quizzes varied slightly from prior experiments in that sometimes unit materials were broken down into smaller subsets for post-lesson quizzing. However, on any given piece of information, students would still be given only one post-lesson quiz. For example, the teacher would teach some lessons on a set of six items, and the students would take a short post-lesson quiz. Then, the teacher would proceed to teach additional content from the same unit (six items), and the students would take another short post-lesson quiz. At the end of the unit, the review quiz would contain all content from previous post-lesson quizzes. Post-lesson and review quizzes were 1 – 13 days apart (M = 6.8 days). The average number of

questions on a post-lesson quiz was six, and the average number of questions on a review quiz was eighteen.

As with Experiment 3, all quiz questions were administered via clicker response system (Ward, 2007). The only procedural difference was that for multiple-choice questions, students were allowed up to thirty seconds to respond. For short-answer items, students were given two minutes to type in their answers.

Paper-and-pencil unit exams were administered by the classroom teacher one or two days after the review quiz. Students were allotted the full class period (approximately 45 minutes) to answer all experimental questions, as well as additional questions written by the teacher and not used in the experiment. Multiple-choice and short-answer questions were presented in a mixed random order on unit exams, and all classroom sections received the same order. Students received their graded unit exams with feedback a few days after completing the unit exams.

Scoring. Short-answer questions required very brief answers, so a simple correct or incorrect coding system was used. The research assistant consulted the classroom teacher with any ambiguities, and they agreed upon the coding (correct or incorrect). An independent research assistant blind to condition also scored a random sample of 10% of the short-answer responses; inter-rater reliability (Cohen's kappa) was .83.

Results

Preliminary considerations. Four students who qualified for special education programs were excluded from the analyses. In addition, 29 students who were not present for all quizzes and exams across both units were excluded from our analyses.

Still, the general pattern of results remained the same with all present and absent students included (see Appendix, Table A-5 for data from all present and absent students). Thus, 40 students contributed data to the present analyses. Once again, means have been collapsed over the two units.

Initial quiz performance. Average performance on the initial quizzes is displayed in Table 5. In general, initial quiz performance increased from the post-lesson quiz (65%, 95% CI [.60, .70]) to the review quiz (70%, 95% CI [.66, .75]). In addition, students answered correctly on the multiple-choice quiz more often than on the short-answer quizzes (83% and 52%, 95% CIs [.80, .86] and [.66, .75], respectively). A 2 (quiz type: post-lesson quiz, review quiz) x 2 (initial quiz format: multiple-choice, short-answer) repeated measures ANOVA confirmed significant main effects of initial quiz format, $F(1, 39) = 108.37$, $p < .001$, $\eta_p^2 = .74$ and quiz type, $F(1, 39) = 9.62$, $p = .004$, $\eta_p^2 = .20$, with no significant interaction, $F(1, 39) = 0.20$, $p = .655$, $\eta_p^2 = .01$. As can be seen in Table 5, students made similar gains in short-answer performance from post-lesson quiz to review quiz as they did for multiple-choice performance (on average, about a 5 percentage point gain for both initial quiz formats, albeit at different points on the performance scales).

Unit exam performance. Average unit exam performance is displayed in Figure 7. In general, students performed best on the unit exam for questions that had occurred on short-answer quizzes (83%, 95% CI [.79, .87]), next best for items that had appeared on the multiple-choice quizzes (81%, 95% CI [.77, .86]), and worst on items not previously tested (69%, 95% CI [.63, .74]), demonstrating once again the large benefits of quizzing on end-of-the-unit retention. A 3 (learning condition: multiple-choice, short-

answer, not tested) x 2 (unit exam format: multiple-choice, short-answer) repeated measures ANOVA revealed significant main effects of learning condition, $F(2, 78) = 25.61$, $p < .001$, $\eta_p^2 = .40$. In addition, there was a significant main effect of exam format, $F(1, 39) = 117.23$, $p < .001$, $\eta_p^2 = .75$, confirming that students answered more multiple-choice items correctly (87%, 95% CI [.83, .90]) than short-answer items (68%, 95% CI [.63, .73]). There was no significant interaction of learning condition and unit exam format, $F(2, 78) = 0.53$, $p = .592$, $\eta_p^2 = .01$. In other words, initial short-answer quizzes and initial multiple-choice quizzes produced similar gains in performance. These results replicate Experiment 1b in that initial quizzes enhanced unit exam performance (i.e. a testing effect was observed). A match between quiz format and unit exam format was not necessary for this benefit, nor did the match enhance the benefit obtained from initial quizzes.

Discussion

The results of Experiment 4 confirm those of Experiment 3 with subjects from high school and with a different subject matter. Once again, quizzing using different question stems produced enhancement on a final unit test. Again, the effect occurred for both multiple-choice and short-answer quizzes and on both types of criterial test. The match between quizzes and criterial test was once again not a factor in the results.

General Discussion

The experiments reported here contain several key findings, which are summarized in Table 6. First, robust quizzing effects were observed in classroom settings, using actual course content in 7th grade science and high school history classes. The use of actual classroom assessments (i.e., unit exams constructed by the

teacher) as the criterial measure is an unusual aspect of the research and is a key feature that makes it directly relevant to classroom settings. The benefits appeared on multiple-choice and short-answer unit exams. These testing effects were consistently observed (see Table 6); across experiments and across conditions, unit exam performance for questions previously quizzed exceeded that for questions not quizzed by 18 percentage points. Quizzing clearly aids performance in classrooms. This outcome replicates prior work (McDaniel et al., 2011, 2012; Roediger et al., 2012), but the current results also extend this work in important ways, delineated below.

First, we consistently observed that the format of the quizzes did not have to match the format of the unit exam for the quizzing benefits to occur. This is the most important finding of the present report in that it is novel, was unanticipated from the laboratory literature, and is a critically important practical point for teachers. Even quick, easily administered multiple-choice quizzes aid student learning, as measured by unit exams (either in multiple-choice or short-answer format).

Further, the benefits were long lasting: Robust effects were seen on the end-of-semester exams in Experiments 1a, 1b, and 2. That is, both multiple-choice and short-answer quizzing enhanced performance on end-of-semester class exams (again, in both multiple-choice and short-answer formats).

A third key finding is that similar benefits were not seen when restudying of target facts was substituted for quizzing (Experiments 2 and 3). That is, low-stakes quizzing in the classroom enhances retention on a delayed test better than does selective restudying of key material, adding to a nascent literature suggesting that retrieval

practice via classroom quizzing does more than simply re-expose the student to materials (Roediger et al., 2011; see also Carpenter et al., 2008).

A fourth finding is that using three initial quizzes (as in many of our prior experiments, e.g., McDaniel et al., 2011) is not necessary to obtain the benefit; Experiments 3 and 4 established these effects with two quizzes.

Changing the wording of the question stem did not eliminate the benefit of quizzing, as seen in Experiments 3 and 4. Students did not simply memorize question-answer pairs without an understanding of the underlying constructs.

The sixth point is that in Experiment 4 the findings were generalized from middle school science classrooms to a high school history class. Importantly, our findings generalize beyond the middle school science classroom.

The effects summarized above were consistent and robust. Other findings were less clear cut, such as the interaction between multiple-choice and short-answer quizzes and final test results. That is, it is not clear from our results whether multiple-choice is a better quiz format than is short-answer (as Experiment 1a might suggest), whether the opposite is true (as the final exam of Experiments 1a and 1b and the unit exam of Experiment 3 might suggest), or whether the two quiz types produce similar benefits (as Experiments 1b and 4 might suggest – as well as our aggregate results across all experiments). What is clear is that both types of quizzes give benefits, and a conservative conclusion across our experiments is that the benefits are roughly equivalent from the two types of test.

Another way of considering the interactions between quiz type and final, criterial test is to consider whether short-answer or multiple-choice criterial exams (i.e., unit or

end-of-semester exams) show differential sensitivity to the format of initial quizzing. To the extent that our results can speak to this issue, they seem to suggest that short-answer criterial exams are more sensitive to prior history of quizzing than are multiple-choice exams. That is, in only one case did we see quiz differences reflected in final multiple-choice performance (i.e., the final exam of Experiments 1a and 1b). In three cases, we observed that final short-answer exams reflected the type of initial quiz taken (i.e., Experiment 1a unit exam, Experiment 1a and 1b final exam, Experiment 3 unit exam). In two of these three cases, a *mismatch* in question type produced the best performance. That is, multiple-choice quizzes led to better performance than did short-answer quizzes on the final short-answer test. This outcome may have occurred because multiple-choice quizzes produced more correct answers (and thus perhaps greater retrieval practice) than did short-answer tests (but of course feedback occurred on all quizzes).

This finding is opposite what one would predict on the basis of the transfer appropriate processing principle (Morris et al., 1977; Roediger et al., 2002). The types of retrieval processes required by multiple-choice quizzes and those needed for short-answer tests differed more (of course) than when the quiz and test format matched, but yet the mismatch condition often led to better performance. The interactions between initial quiz type and final, criterial test type were varied enough in these experiments that a better understanding of what pattern happens when awaits future work. The advantages of quizzing are quite clear, however, and when differences were seen in quiz format, they were small in comparison to the clear advantages seen overall.

These findings are especially noteworthy in that they occurred in authentic middle-school and high-school classrooms with normal, closed-book quizzes (and not open-book quizzes used with online college classes, as examined by McDaniel et al., 2012). The spacing between quizzes and between the final quiz and the unit exam was always at least one day and occurred with a natural spacing, suitable for typical student classrooms.

It remains for future work to discover whether repeated quizzes are necessary for these beneficial effects or whether a single quiz might be sufficient. Laboratory work would suggest that more tests would be better (e.g., McDermott, 2006), but as the present studies demonstrate, predictions derived from laboratory studies do not always hold up within classroom settings (see McDaniel, 2011 on this point). That is, both laboratory and theoretical principles would have led to the prediction that benefits from short-answer quizzing would exceed those from multiple-choice quizzing (e.g., Kang et al., 2007). In this way, the present studies also point to the importance of field studies for establishing best-practice recommendations for school classrooms.

One possible concern regards spillover effects from experiment-to-experiment conducted on the same students in the same academic year. The dates of testing are shown in Table 6. Experiments 1a, 1b, and 2 occurred with the same students, back-to-back within one school year. Experiment 3 took place in the same classroom but with a new set of students (in a new school year). Experiment 4 involved a different classroom and a different set of (older) students. Hence, the spillover concerns apply only to Experiments 1a, 1b, and 2. If such effects occurred and influenced the results, we might expect that student performance would increase over time during the year and that the

benefits of quizzing might wane (if students began to administer self-quizzing to all materials). That is, by Spring of the school year, when Experiment 2 was administered, students could have used self-testing for the restudied items to achieve the benefits of quizzing. The benefits of quizzing (over restudying) remained even under these conditions, however. Of course, any intervention within a classroom to use quizzing to enhance learning would occur under similar circumstances. Spillover effects do not seem to be a problem, which may indicate that middle school students do not appreciate the benefits of quizzing and use it as a study strategy even when it has been shown to work for them (the same holds true in college students; see Karpicke, 2009).

One limitation of the current work is that the learning conditions were intermixed; a multiple-choice question could precede a short-answer question, followed by another multiple-choice question. As such, we cannot say for sure that our results would generalize to a situation in which the quiz questions are blocked by type (or manipulated between-students). If, for example, students adopted a short-answer question mindset – trying to generate an answer even in multiple-choice tests before examining the choices -- when approaching all the quiz questions in the present studies, this could account for the similarities in quiz type. What we can say for sure is that when quiz questions are intermixed, question type does not exert much influence.

From a practical standpoint, our findings point the way to recommendations for class settings. Specifically, exactly what format is used for the quizzes and when the quizzes are given are not major determinants of the degree of benefit from quizzing. Clearly, in-class quizzes with feedback enhance student performance on later summative exams whether they are short-answer or multiple-choice.

This unanticipated finding is even more important in light of the consideration that multiple-choice quizzes take much less class time and less teacher involvement (in terms of grading) than do short-answer quizzes. In the present experiments, short-answer tests took longer than did multiple-choice tests; time-on-task was not equated (and indeed differed substantially). Note, though, that in Experiments 1a and 1b, the multiple-choice tests led to greater performance than did the short-answer test, despite the fact that the short-answer tests took 2.5 times as long to administer. From an efficiency point of view, classroom time would be optimized by administering multiple-choice quizzes, which give robust benefits to learning while taking minimal class time and requiring less of the teacher's time for grading. It is also worth noting that retrieval practice effects occurred both with a clicker system and with paper-and-pencil exams. Special equipment is not necessary to obtain these quizzing benefits.

In summary, cross-format benefits from in-class quizzes were seen within middle school science classroom, using class materials and standard assessments constructed by the teachers and taken for a grade. This benefit adds to the growing literature pointing to the utility of low-stakes testing in enhancing student learning.

References

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.

Agarwal, P. K., Bain, P. M., & Chamberlain, R. C. (2012). The value of applied research: Retrieval practice improves classroom learning and recommendations from a teacher, a principal, and a scientist. *Educational Psychology Review*, *24*, 353-354.

Arnold, K.M. & McDermott, K.B. (2013a). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 940-945.

Arnold, K.M. & McDermott, K.B. (2013b). Free recall enhances subsequent learning. *Psychonomic Bulletin & Review, 20,* 507-513.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185-205). Cambridge, MA: MIT Press.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56-64). New York: Worth Publishers.

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.

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.

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

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.

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

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

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.

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.

Karpicke, J.D. (2009). Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General, 138*, 469-486.

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

McDaniel, M. A., Roediger, H. L., & 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., McDermott, K.B. & Roediger, H.L. (2013). Quizzing in middle school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology, 27,* 360-372.

McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study. *Journal of Applied Research in Memory and Cognition*, 1, 18-26.

McDermott, K.B., Arnold, K.M., & Nelson, S.M. (in press). The Testing Effect. T. Perfect and S. Lindsay (Eds). *Sage Handbook of Applied Memory.* Sage.

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

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied, 17,* 382-395.

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.

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

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

Ward, D. (2007). eInstruction: Classroom Performance System [computer software]. Texas: eInstruction Corporation.

Table 1. Average initial quiz performance (proportion correct) as a function of quiz placement and question format. Data are from Experiment 1a. Standard errors are shown in parentheses.

	Multiple-choice quiz	Short-answer quiz	Overall
Pre-lesson quiz	.37 (.03)	.15 (.02)	.26 (.02)
Post-lesson quiz	.73 (.03)	.43 (.03)	.58 (.02)
Review quiz	.87 (.01)	.63 (.03)	.75 (.02)
Overall	.66 (.02)	.40 (.02)	

Table 2. Average initial quiz performance (proportion correct) as a function of quiz placement and question format. Data are from Experiment 1b. Standard errors are shown in parentheses.

	Multiple-choice quiz	Short-answer quiz	Overall
Post-lesson quiz 1	.69 (.02)	.23 (.02)	.46 (.02)
Post-lesson quiz 2	.79 (.02)	.37 (.02)	.58 (.02)
Review quiz	.86 (.01)	.55 (.03)	.71 (.02)
Overall	.78 (.02)	.38 (.02)	

Table 3. Average initial quiz performance (proportion correct) in Experiment 2. Standard errors are shown in parentheses.

	Short-answer quiz
Post-lesson quiz 1	.42 (.02)
Post-lesson quiz 2	.59 (.02)
Review quiz	.74 (.02)

Table 4. Average initial quiz performance (proportion correct) as a function of quiz placement and question format. Data are from Experiment 3. Standard errors are shown in parentheses.

	Multiple-choice quiz	Short-answer quiz	Overall
Post-lesson quiz	.72 (.02)	.40 (.02)	.56 (.01)
Review quiz	.83 (.01)	.56 (.02)	.70 (.02)
Overall	.78 (.01)	.48 (.02)	

Table 5. Average initial quiz performance (proportion correct) as a function of quiz placement and question format. Data are from Experiment 4. Standard errors are shown in parentheses.

	Multiple-choice quiz	Short-answer quiz	Overall
Post-lesson quiz	.81 (.02)	.49 (.04)	.65 (.02)
Review quiz	.85 (.01)	.55 (.04)	.70 (.02)
Overall	.83 (.02)	.52 (.03)	

Table 6. Summary of results across Experiments 1-4. Robust quizzing benefits were seen in all experiments.

Experiment and Dates of data collection	Grade	Topic	Number of quizzes	Wording	Design	Testing effect on unit exam?	Unit exam depend on quiz type?	Restudying as good as quizzing?	Quiz form (clicker/paper and pencil)	Primary conclusion
1a F09	7	science	3	identical	3 (learning condition: MC, SA, NT) x 2 Final Test (MC, SA)	Yes 74% > 64%	Yes. SA exam: MC quizzes better than SA	n/a	MC: clicker SA: paper & pencil	A match in test format is not necessary to achieve benefits of retrieval practice.
1b F09-S10	7	science	3	identical	3 (learning condition: MC, SA, NT) x 2 Final Test (MC, SA)	Yes 73% > 55%	No	n/a	MC: clicker SA: paper & pencil	A match in test format is not necessary to achieve benefits of retrieval practice.
2 S10	7	science	3	identical	3 (learning condition: SA, restudy, NT). Final Test SA	Yes 81% > 55%	n/a	No	SA only (paper & pencil)	Retrieval practice benefits go beyond selective re-exposure of the information.
3 F11	7	science	2	changed	4 (learning condition: MC, SA, restudy, NT) x 2 Final Test (MC, SA)	Yes 84% > 72%	Yes. SA exam: SA quizzes better than MC	No	MC: clicker SA: clicker	Only 2 quizzes may be sufficient. The questions do not need to be identical.
4 F11	11,12	history	2	changed	3 (learning condition: MC, SA, NT) x 2 Final Test (MC, SA)	Yes 82% > 69%	No	n/a	MC: clicker SA: clicker	Effects generalize beyond middle school science. Again, changing

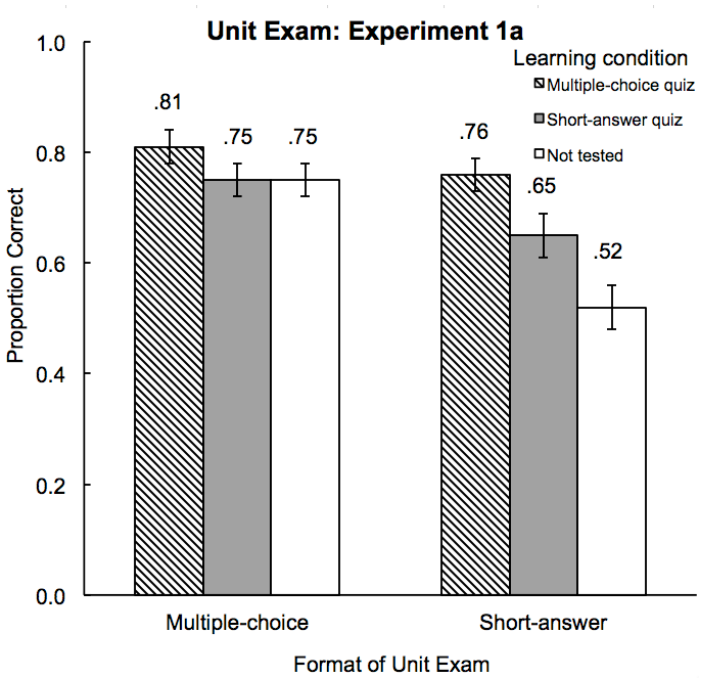

Figure 1. Average unit exam performance (proportion correct) as a function of learr condition and unit exam format. Data are from Experiment 1a. Error bars represent standard errors of the mean.

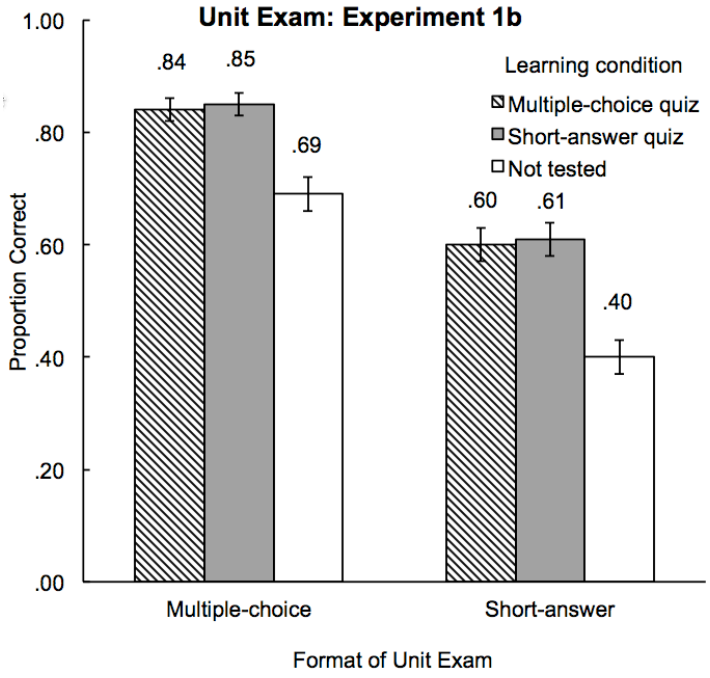

Figure 2. Average unit exam performance (proportion correct) as a function of learning condition and unit exam format. Data are from Experiment 1b. Error bars represent standard errors of the mean.

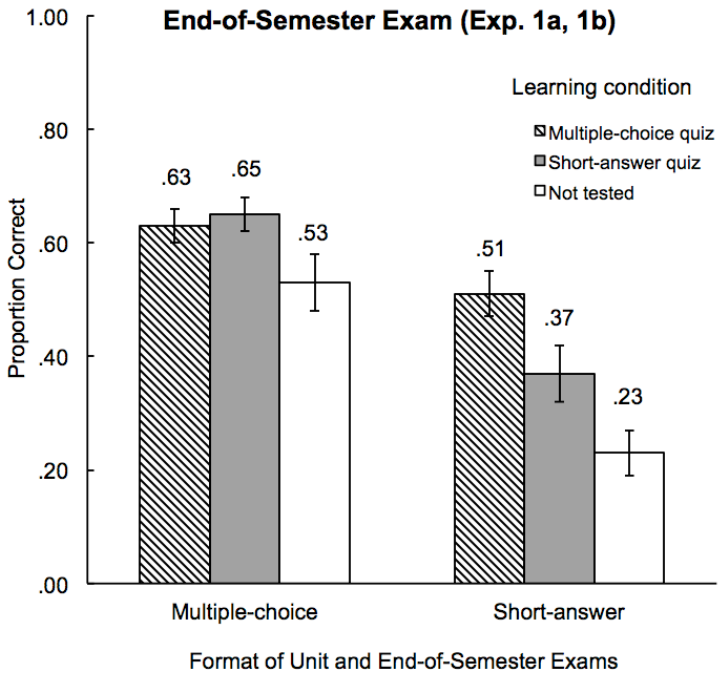

Figure 3. Average end-of-the-semester exam performance (proportion correct) as a function of learning condition and unit exam format. Data are collapsed over Experiments 1a and 1b. Error bars represent standard errors of the mean.

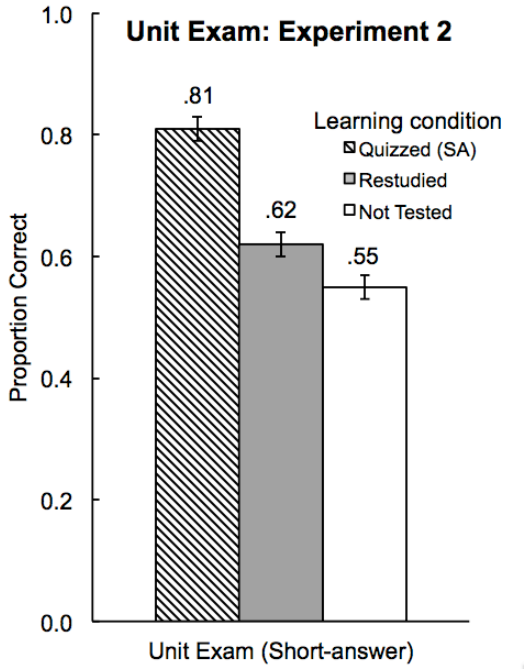

Figure 4. Average unit exam performance (proportion correct) as a function of learning condition. Data are from Experiment 2. Error bars represent standard errors of the mean.

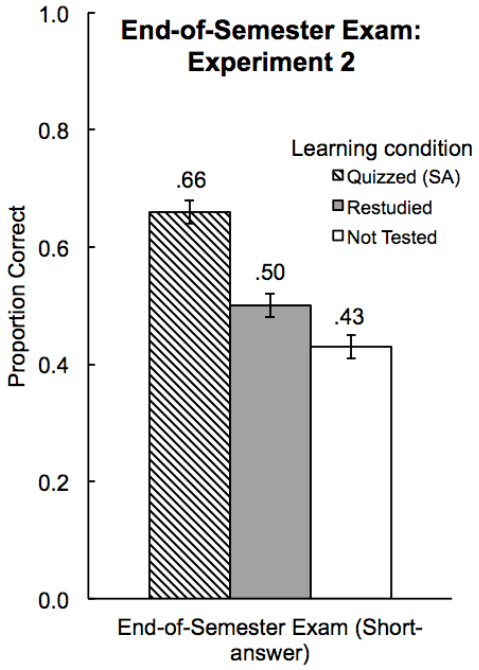

Figure 5. Average end-of-the-semester exam performance (proportion correct) as a function of learning condition. Data are from Experiment 2. Error bars represent standard errors of the mean.

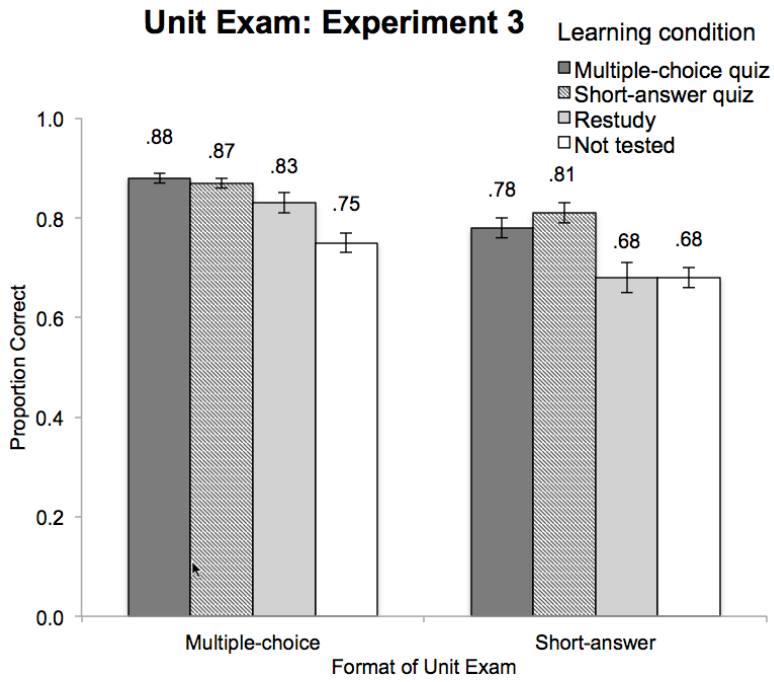

Figure 6. Average unit exam performance (proportion correct) as a function of learning condition and unit exam format. Data are from Experiment 3. Error bars represent standard errors of the mean.

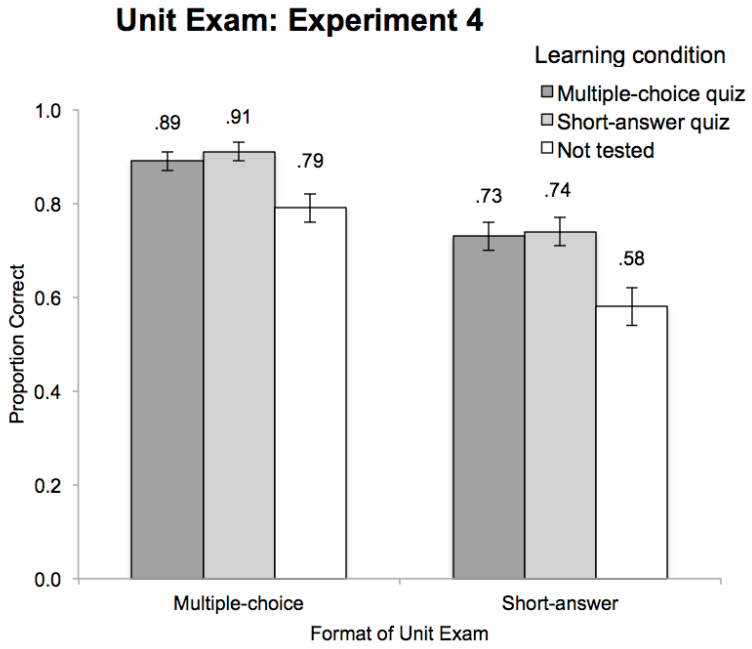

Figure 7. Average unit exam performance (proportion correct) as a function of learning condition and unit exam format. Data are from Experiment 4. Error bars represent standard errors of the mean.

Appendix

Examples of Learning Conditions

Learning condition	Example
Multiple-choice quiz	What are two characteristics of angiosperms? A. They produce needles and they have naked seeds. B. They have five petals and two cotyledons. C. They produce flowers and fruits. D. They produce pollen and cones.
Short-answer quiz	What are two characteristics of angiosperms? _____ _____ _____
Restudy	Two characteristics of angiosperms are that they produce flowers and fruits.

Examples of Initial Quiz Formats for Experiment 3

Type of Test	Type of Question	Example
Post	Multiple-choice	Hundreds of years ago, when people believed life could appear suddenly from nonliving things, their mistaken idea was known as _____. A. evolution B. spontaneous combustion C. spontaneous generation D. transformation
Review	Multiple-choice	What is the idea that living things arise from nonliving sources, such as flies arising from decaying meat? A. random birth B. spontaneous generation C. symbiotic functioning D. sporadic construction
Unit Exam	Multiple-choice	When frogs appeared in mud puddles after heavy rains, people concluded frogs could sprout from mud in ponds. What is the term for their mistaken belief that life could come from nonliving sources? A. spontaneous generation B. random assignment C. the big bang theory D. evolution
Post	Short-answer	Hundreds of years ago, people believed life could appear suddenly from nonliving things. What was their mistaken idea known as?
Review	Short-answer	What is the idea that living things arise from nonliving sources, such as flies arising from decaying meat?
Unit Exam	Short-answer	When frogs appeared in mud puddles after heavy rains, people concluded frogs could sprout from mud in ponds. What is the term for their mistaken belief that life could come from nonliving sources?
Post	Restudy	Hundreds of years ago, when people believed life could appear suddenly from nonliving things, their mistaken idea was known as spontaneous generation.
Review	Restudy	The idea that living things arise from nonliving sources, such as flies arising from decaying meat is known as spontaneous generation.

Examples of Initial Quiz Formats for Experiment 4

Type of Test	Type of Question	Example
Post	Multiple-choice	What act of congress prevented discrimination in employment and public accommodations, and provided the federal government with the powers to enforce desegregation? A. 13th Amendment B. Equal Rights Act of 1964 C. Civil Rights Act of 1964 D. Affirmative Action
Review	Multiple-choice	Because it involves discrimination in employment, refusing to hire someone solely because they are African American would be a violation of which act of congress? A. Voting Rights Act B. Civil Rights Act of 1964 C. Civil Rights Act of 1970 D. All of the above
Unit Exam	Multiple-choice	Because it involves discrimination in public accommodations, requiring African American people to use a separate water fountain at a state park would be a violation of which act of congress? A. Brown vs. Board of Education B. Civil Rights Act of 1964 C. 15th Amendment D. All of the above
Post	Short-answer	What act of congress prevented discrimination in employment and public accommodations, and provided the federal government with the powers to enforce desegregation?
Review	Short-answer	Because it involves discrimination in employment, refusing to hire someone solely because they are African American would be a violation of which act of congress?
Unit Exam	Short-answer	Because it involves discrimination in public accommodations, requiring African American people to use a separate water fountain at a state park would be a violation of which act of congress?

Table A-1. Average unit exam performance as a function of learning condition and unit exam format. Data are from Experiment 1a, including present and absent students (N = 106). Standard errors are shown in parentheses.

	Unit Exam Format	
Learning condition	Multiple-choice	Short-answer
Multiple-choice quiz	.79 (.02)	.70 (.03)
Short-answer quiz	.73 (.02)	.59 (.03)
Not tested	.72 (.02)	.52 (.03)

Table A-2. Average unit exam performance as a function of learning condition and unit exam format. Data are from Experiment 1b, including present and absent students (N = 106). Standard errors are shown in parentheses.

	Unit Exam Format	
Learning condition	Multiple-choice	Short-answer
Multiple-choice quiz	.77 (.02)	.59 (.02)
Short-answer quiz	.79 (.02)	.55 (.02)
Not tested	.66 (.02)	.37 (.02)

Table A-3. Average unit and delayed exam performance as a function of learning condition. Data are from Experiment 2, including present and absent students (N = 106). Standard errors are shown in parentheses.

Learning condition	Unit Exam	Delayed Exam
Quizzed	.77 (.02)	.64 (.02)
Restudied	.58 (.02)	.48 (.02)
Not Tested	.51 (.02)	.43 (.02)

Table A-4. Average unit exam performance as a function of learning condition and unit exam format. Data are from Experiment 3, including present and absent students (N = 116). Standard errors are shown in parentheses.

	Unit Exam Format	
Learning condition	Multiple-choice	Short-answer
Multiple-choice quiz	.85 (.01)	.74 (.02)
Short-answer quiz	.86 (.01)	.78 (.02)
Restudied	.80 (.01)	.65 (.02)
Not tested	.75 (.02)	.62 (.02)

Table A-5. Average unit exam performance as a function of learning condition and unit exam format. Data are from Experiment 4, including present and absent students (N = 69). Standard errors are shown in parentheses.

	Unit Exam Format	
Learning condition	Multiple-choice	Short-answer
Multiple-choice quiz	.88 (.02)	.72 (.03)
Short-answer quiz	.89 (.02)	.71 (.02)
Not tested	.80 (.02)	.57 (.03)