Journal of Educational Psychology
2011, Vol. 103, No. 2, 399–414

0022-0663/11/$12.00 DOI: 10.1037/a0021782

Test-Enhanced Learning in a Middle School Science Classroom: The Effects of Quiz Frequency and Placement

Mark A. McDaniel and Pooja K. Agarwal
Washington University in St. Louis

Barbie J. Huelser
Columbia University

Kathleen B. McDermott and Henry L. Roediger, III
Washington University in St. Louis

Typically, teachers use tests to evaluate students' knowledge acquisition. In a novel experimental study, we examined whether low-stakes testing (*quizzing*) can be used to foster students' learning of course content in 8th grade science classes. Students received multiple-choice quizzes (with feedback); in the quizzes, some target content that would be included on the class summative assessments was tested, and some of the target content was not tested. In Experiment 1, three quizzes on the content were spaced across the coverage of a unit. Quizzing produced significant learning benefits, with between 13% and 25% gains in performance on summative unit examinations. In Experiments 2a and 2b, we manipulated the placement of the quizzing, with students being quizzed on some content prior to the lecture, quizzed on some immediately after the lecture, and quizzed on some as a review prior to the unit exam. Review quizzing produced the greatest increases in exam performance, and these increases were only slightly augmented when the items had appeared on previous quizzes. The benefits of quizzing (relative to not quizzing) persisted on cumulative semester and end-of-year exams. We suggest that the present effects reflect benefits accruing to retrieval practice, benefits that are well established in the basic literature.

Keywords: test-enhanced learning, quiz-enhanced learning, quiz frequency, quiz placement, middle school science

Basic memory experiments have shown that on a final criterial test, students better remember information on which they had been tested sometime prior to the test than information on which they had not been tested previously (see Roediger & Karpicke, 2006, for a review). This effect, termed the *testing effect*, suggests an important expansion of how tests might be utilized in educational settings. Currently, tests largely serve an evaluative function to help teachers to gauge students' knowledge acquisition and achievement and to assign grades. In the present article, we explore the idea that low- or no-stakes testing (i.e., *quizzing*) might be used as a technique to improve learning and retention of course content in a middle school classroom.

The idea that testing can be used to enhance learning and retention has been explored in a handful of experimental studies reported in the educational psychology literature. For instance, Spitzer (1939) presented thousands of Iowa middle school students with a passage to read and after delays of varying length gave the students a test on the material. Final test performance was better when intervening tests were required than when no intervening tests were present (for related research with college students, see Glover, 1989; McDaniel & Fisher, 1991). These and other reports of the testing effect with educational-like materials (see Bangert-Drowns, Kulik, & Kulik, 1991; Roediger, Agarwal, Kang, & Marsh, 2010, for reviews) underscore the potential for testing as a classroom technique to enhance learning. Yet features inherent in much of the extant research on testing effect are not reflective of acquisition of curricular material in a classroom. For example in Spitzer (1939), the testing effect was demonstrated for material that students were exposed to once and to which students had no further access for review and study. Further, the material was an isolated passage not related to the integrated content representing the educational objectives of the class. By contrast, in a classroom context. material is typically reinforced in homework and reading

This article was published Online First February 21, 2011.

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

This research was supported by Grant R305H060080-06 awarded to Washington University in St. Louis from the U.S. Department of Education, Institute of Education Sciences. 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 and to Leo Sherman and Jack Turner, who were the superintendents during the collection of these data. We especially thank Roger Chamberlain, the Columbia Middle School principal; Ammie Koch, the science teacher; and all of the 2007–2009 eighth grade students and parents. We also thank Lindsay Brockmeier, Jane McConnell, Kari Farmer, and Jeff Foster for their assistance.

Correspondence concerning this article should be addressed to Mark A. McDaniel, Washington University in St. Louis, Department of Psychology, Campus Box 1125, St. Louis, MO 63130-4899. E-mail: mmcdanie@artsci.wustl.edu

assignments, it is designated as important for the students to master, and the material is part of an integrated topic domain identified as core to the curriculum.

We are aware of only a few published experiments in which the testing effect was investigated for content presented in an actual course. One experiment was performed in a college-level web-based course (McDaniel, Anderson, Derbish, & Morrisette, 2007), and the other was associated with an eighth-grade U.S. history class (Carpenter, Pashler, & Cepeda, 2009). In both cases, information tested on initial short-answer tests was more likely to be remembered than information not included in the initial test or information presented for additional review. One limitation of these studies, however, is that the final criterial tests were not the summative tests used to evaluate students for the course. For the McDaniel et al. experiment, the final tests for the content were termed "practice" tests (tests that did not affect students' grades and were optional); for the Carpenter et al. experiment, the initial and final tests were administered after the students had completed their examinations (thus, students' grades were established prior to their participation in the experiment), and students were unaware that a final test would be administered. Accordingly, students were not as likely to be motivated to study the target material in preparation for the criterial tests or to learn the target material in the first place (in McDaniel et al.) than if their course grade had depended on their test performance. It is thus possible that testing (quizzing) might not produce significant benefits to learning and retention in the authentic classroom context in which performance on the criterial tests is important for the course grade, and students are motivated to study the target content to do well on the final assessments. In this context, quizzed and unquizzed content might be equally well learned. Thus, the question that remains is whether low-stakes quizzing could be used to promote learning for core curricular content in K–12 (or college) classrooms (e.g., see Mayer et al., 2009, for a quasi-experiment in college classes in which no effects on course grades were found when instructors administered between two and four multiple-choice questions at the end of lectures relative to a no-quiz class).

There are, however, a number of theoretical reasons why quizzing should promote learning. Quizzing requires active processing of the target material and more specifically requires retrieval, a process that improves retention (Carpenter & DeLosh, 2006; McDaniel & Masson, 1985; Roediger & Karpicke, 2006). Quizzing is usually accompanied by feedback (as in the current study), which itself improves learning (Butler & Roediger, 2008; Pashler, Cepeda, Wixted, & Rohrer, 2005). Quizzing could also have indirect effects, such as improving students' metacognitive judgments about what they know and do not know (Kornell & Son, 2009), thereby increasing study effectiveness (Thomas & McDaniel, 2007). Frequent quizzing might also reduce test anxiety, thereby improving performance on summative assessments. The experiments reported here are part of an ongoing comprehensive project to evaluate whether classroom quizzes presented as learning exercises, on which performance has little consequence for the students' grades, will indeed promote learning across a range of middle school courses.

The present emphasis on quizzing is not intended to imply that other kinds of activities, especially those that include feedback, such as in-class reviews, homework, self-explanation (e.g., McDaniel & Donnelly, 1996), and open-book questions (e.g., Agarwal, Karpicke, Kang, Roediger & McDermott, 2008; Callender & McDaniel, 2007), are not also effective active learning techniques. Our emphasis is motivated by the observation that quizzing is not often incorporated into the arsenal of techniques that teachers employ but may be an efficacious technique. In our initial work, we focused on middle school social studies classes in which the instructor had already adopted quizzing (one of the few in the school) to assist students in learning and not for grading purposes. She had established a particular quizzing regimen in her classes in which students received a prelecture quiz, postlecture quiz, and a review quiz, all of which were identical in content. In several experiments, we found that this particular quizzing regimen significantly enhanced performance on the summative assessments for items that were quizzed relative to items that were either not quizzed or were presented for restudy instead of as a quiz item (Roediger, Agarwal, McDaniel, & McDermott, 2010).

Several important questions emerged from these initial findings (conducted in parallel with the current study). First, could similar positive effects of quizzes be obtained for middle school science? Improving science education has become a national priority in terms of policy, funding, and educational research. If effective, quizzing could be an attractive tool in efforts to enhance students' science literacy for a number of reasons, including minimal disruption to existing classroom practice. Accordingly, the present experiments were conducted across a range of eighth-grade science content. A second key question was whether the significant benefits of quizzing rest on the particular three-quiz regimen reported by Roediger, Agarwal, McDaniel, et al. (2010). From a practical standpoint, educators would likely prefer to implement the most efficacious quizzing scheme. To provide information along these lines, in Experiments 2a and 2b we systematically manipulated the number of quizzes and their placement relative to the classroom lesson and the summative assessment.

Third, the robustness of quizzing effects (if found) for long retention intervals has received little attention. A recent laboratory experiment showed testing effects for art-history content after a 2-month delay (Butler & Roediger, 2007). The Carpenter et al. (2009) study showed testing effects after a 9-month delay for eighth graders but only for facts tested once the lessons and exams were completed. Thus, though these prior studies established that testing effects can persist after a substantial delay for educational material, in our current experiments we examined students' learning of course material as they progressed through the course, and the criterial tests we used were the chapter tests and exams on which students were graded. As far as we can tell, no previous researchers have conducted experiments integrated into the subject matter of a class in this way. Quizzing would be an especially valuable pedagogical technique if its use during a course supported long-term retention of course content. To examine the persistence of quizzing benefits in a classroom, in Experiments 1 and 2b we examined student performances on end-of-the semester and end-of-the year cumulative exams.

Experiment 1

In a within-student design, all students received three multiple-choice quizzes (with feedback); for each quiz some of the target content (i.e., content that would be included on the class exams) was included and some of the target content was not included, and

across six class sections the particular content for quizzing (or not quizzing) was randomly determined. Quizzed content was quizzed prelecture, immediately postlecture, and a day prior to the unit exam. Performance on the class's unit and cumulative examinations (containing both quizzed and nonquizzed items) indicated whether quizzing affected learning and retention.

Method

Participants. We recruited 139 eighth-grade science students from a public middle school in a suburban middle-class community in the Midwest to participate in this study. Parents were informed of the study, and written assent from each participant was obtained in accordance with the Washington University Human Research Protection Office. The school board, the principal, and the teacher agreed to participate in the study; three students declined to have their data included in the study.

Materials and design. We used material from five units in the assigned science curriculum (genetics, evolution, anatomy 1, anatomy 2, and anatomy 3). There were three initial quiz phases: prelesson (before the teacher's lesson but after participants read an assigned chapter from the textbook [assuming students followed the teacher's instructions]), postlesson (after the teacher's lesson about a chapter), and review (24 hr before the unit exam). On the initial classroom quizzes, half of the target facts from each unit appeared on the test in a multiple-choice format (quizzed condition) and half of the facts did not (nonquizzed condition), following a within-subjects design. For these initial quizzes (not the review quiz), the number of questions varied across each unit, depending on the length of the chapters. The number of questions on the initial quiz (i.e., the prelesson and postlesson quizzes) ranged from three for the shortest chapter to eight questions for the longer chapters. All facts were covered in the readings and the teacher's lessons.

The classroom teacher approved all multiple-choice questions, answers, and lures created by the experimenter. Most of the questions were based on examinations that this teacher had used in previous years, and thus the questions were reflective of the kind of summative assessments routinely used for eighth-grade science at this school. These assessments primarily were based on factual multiple-choice questions; additional multiple-choice questions created by the experimenter typically required inference and analysis (across the five units, 80% of the questions were factual and 20% required inference or analysis; see Appendix A for example items). Items were randomly assigned to the two conditions, and each of the six classroom sections ($M = 24$ students) received a different random selection of items in the quizzed and nonquizzed conditions. The number of items varied between conditions and units (ranging from 12 to 30 items per condition and unit), and the total number of items in this experiment was 188. For each student, 96 items were in the quizzed condition and 92 items were in the nonquizzed condition.

Twenty days (on average) after the first prelesson quiz, retention was measured on unit exams composed of all items noted previously. In addition, the unit exams included a section with various types of other questions that the teacher generated (matching, fill in the blank, short answer). The exact nature of these items varied between each unit. For example, in genetics, the students had to fill out Punnett squares and answer questions regarding the phenotype and genotype of the offspring; for anatomy, students had to label parts of the systems they were learning. Depending on the unit, these additional, nonmultiple-choice items represented 15%–56% of the questions on the exams (40% averaged across exams); further, about 5% of the questions on the exams were multiple-choice items that were added by the teacher subsequent to the lessons and consequently were not in the pool of items on the quizzes. Performance on these additional items was not examined for the current study (because the quiz manipulations were within-subjects—all students took quizzes—the effects of quizzing could not be determined on these questions). It is worth emphasizing, however, that the multiple-choice examination questions analyzed for the current experiments were for the most part those used by the teacher and had typically been used in previous years to evaluate and grade the eighth-grade science students.

Procedure. Initial quizzes (prelesson, postlesson, and review) were administered via a clicker response system (Ward, 2007). Items on initial quizzes were presented in the same order as presented during the lessons. Order of multiple-choice alternatives was randomized for each quiz to prevent students from memorizing the location of the correct answer.

Prelesson quizzes were administered after students read an assigned chapter from the textbook but before the teacher discussed the information. The teacher was not present for these quizzes in order to prevent potential bias toward particular items during her lesson that immediately followed the prelesson quiz. Students were truthfully informed that the teacher had to leave the room so that she would not know which questions were on the quiz; otherwise the results could be "messed up." Students were encouraged to pay attention to the quiz questions because the information would likely be in the lecture and might be on later tests. If students inquired whether the quiz questions would be on the exam, the experimenter responded, "I do not know exactly which questions will be on the exam. Everything is randomized."

For each item, the question and four multiple-choice alternatives were displayed on a large projection screen at the front of the classroom while they were read aloud by the experimenter. Students were required to respond to each question by pressing the A, B, C, or D buttons on their individual clicker remotes. After all the students had responded, a green check mark appeared next to the correct response while the experimenter read the question stem and correct answer out loud to the class before proceeding to the next item. After the completion of the prelesson quiz, the teacher was brought back into the room, and anonymous scores of all students were shown briefly on the screen. Students knew their *own* score by their assigned clicker number. The teacher then proceeded with the lesson.

Postlesson quizzes were administered after the teacher covered all material for a particular chapter. Review quizzes were administered 24 hr before unit exams. Overall, the procedure for postlesson and review quizzes was identical to that used for prelesson quizzes with two exceptions: the teacher was present during these quizzes, and scores from these quizzes counted for a small portion (10%) of each student's cumulative grade. Additionally, students were not explicitly told when postlesson quizzes would be administered, but they were aware that the review quiz was the day before the unit exam. After the review quiz, students were reminded that many questions would be on the unit exam that students had not previously seen.

The classroom teacher administered paper-and-pencil unit exams. The students had been quizzed previously on approximately half of the target items on the exam and had not been quizzed on the other half. For previously quizzed items, the multiple-choice questions on the unit exams were the same as those on the initial quizzes, but the four multiple-choice alternatives were reordered randomly. The classroom teacher used the scores on these exams to account for 50% of the students' overall grade. The students were informed of their overall score the day after the unit exam, but they did not receive corrective feedback on an item-by-item basis.

Students also completed multiple-choice end-of-the-semester and end-of-the-year exams (the end-of-year exams were administered via the clicker response system). On each exam, half the facts had previously appeared on quizzes three times and half had not (note that for the end-of-semester exam, but not the end-of-year exam, there were also multiple-choice questions that targeted units not included in the current experiment). All facts had been tested once on the unit exam, but items on the end-of-the-year exam were not presented on the end-of-the-semester exam. The end-of-the-semester exam was composed of 100 target items (20 items per each of the five units, 10 had been on previous quizzes and 10 had not) and additional items from units not involved in the experiment; the end-of-the-year exam was composed of 10 target facts total (two from each of the five units, one that had appeared on quizzes and one that had not). Questions were presented in the order in which the units had been taught, and questions for each unit were presented in a different random order for each classroom section. Regarding the end-of-the-semester exam, students were notified of the date of the exam approximately 1 month before it was administered, as performance was recorded for grading purposes (20% of their cumulative grade). The retention intervals between the unit exams and the end-of-the-semester exam ranged from 3 months (for the first unit in the semester—genetics) to several days (for the last unit of the semester—astronomy 3). Regarding the end-of-the-year exam, students were not informed of the exam until it was administered, and performance on the exam did not count toward the students' grades. The retention intervals between the unit exams and the end-of-the-year exam ranged from 5 months to approximately 8 months.

The teacher's typical lesson plans remained unchanged throughout our procedure. Students were exposed to all of the information contained on the unit exam via the teacher's lessons, homework, and worksheets; therefore, students were exposed at least once to items that had not appeared on quizzes during typical classroom activities.

Results

Nineteen students who qualified for special education or gifted programs were excluded from our analyses. Furthermore, 28 students who were not present for all initial quizzes, unit exams, and delayed exams were also excluded to ensure the integrity of our quizzing schedule. Therefore, data from 92 students are reported below. However, the general pattern of results remained the same when all 139 students were included. All results were significant at an alpha level of .05 unless otherwise noted. To index effect sizes, for the F tests, partial eta-squared values were computed and for the t tests, Cohen's d values were computed.

Initial quiz performance. Initial quiz performance as a function of unit and type of quiz is shown in Table 1. A 5 (unit) × 3 (quiz type: prelesson, postlesson, review) repeated-measures analysis of variance (ANOVA) confirmed a significant increase from the prelesson (58%) to the postlesson (83%) and review quizzes (86%), $F(2, 182) = 1183.38$, $\eta_p^2 = .93$ for the main effect. Pairwise comparisons indicated that postlesson performance and review quiz performance were significantly greater than prelesson quiz performance, $t(91) = 38.73$, $d = 2.52$, and $t(91) = 38.95$, $d = 2.99$, respectively. Review quiz performance was also significantly greater than postlesson performance, $t(91) = 7.03$, $d = 0.45$. These results demonstrate substantial learning from the teacher's lesson between prelesson and postlesson quizzes, with an additional increase in learning from the postlesson to the review quiz. Performance also differed depending on the unit of material, $F(4, 364) = 12.74$, $\eta_p^2 = .12$ for the main effect. Finally, though review quiz performance was typically greater than postlesson performance, which was always greater than prelesson performance, these differences varied as a function of the unit, $F(8, 728) = 32.60$, $\eta_p^2 = .26$, for the interaction. For instance, several units showed learning primarily between the prelesson and postlesson quizzes, whereas other units (evolution and anatomy 3) also showed learning between the postlesson and review quizzes.

Unit exam performance. Unit exam performance as a function of unit and learning condition (quizzed, nonquizzed) is shown in Table 2. A 5 (unit) × 2 (quizzed, nonquizzed) ANOVA showed that students' unit exam performance on quizzed items (92%) was significantly greater than performance on nonquizzed items (79%), $F(1, 91) = 337.99$, $\eta_p^2 = .79$. There was no significant difference in students' performance across the units, $F(4, 364) = 1.44$, $p > .05$; however, students' relative performance on the quizzed and nonquizzed items varied as a function of the unit, $F(4, 364) = 4.60$, $\eta_p^2 = .05$. As can be seen in Table 2, the testing effect (difference between items on which the students had been quizzed and items on which they had not) ranged from 16% for the genetics

Table 1
Students' Average Initial Quiz Performance as a Function of Unit and Type of Quiz in Experiment 1

	Genetics		Evolution		Anatomy 1		Anatomy 2		Anatomy 3		Overall	
Quiz	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Prelesson	.68	.14	.59	.15	.61	.15	.46	.17	.58	.15	.58	.11
Postlesson	.85	.11	.83	.12	.81	.12	.87	.12	.78	.14	.83	.08
Review	.86	.11	.89	.10	.83	.09	.86	.13	.88	.13	.86	.08

Table 2
Students' Average Unit and Delayed Examination Performance as a Function of Unit and Learning Condition in Experiment 1

	Genetics		Evolution		Anatomy 1		Anatomy 2		Anatomy 3		Overall	
Examination/content	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Unit												
Quizzed	.92	.09	.93	.07	.92	.08	.93	.11	.91	.10	.92	.06
Nonquizzed	.76	.13	.78	.15	.81	.10	.80	.16	.82	.15	.79	.10
End-of-the-semester												
Quizzed	.75	.19	.81	.19	.71	.19	.81	.20	.87	.13	.79	.12
Nonquizzed	.70	.20	.71	.21	.64	.20	.76	.17	.79	.17	.72	.13
End-of-the-year												
Quizzed	.61	.49	.53	.50	.60	.49	.89	.31	.77	.42	.68	.21
Nonquizzed	.63	.49	.45	.50	.59	.50	.73	.45	.73	.45	.62	.22

unit to 9% for the anatomy unit, but all testing effects for each unit were significant, ps $< .05$.

Delayed exam performance. Performances on end-of-the-semester and end-of-the-year exams are also displayed in Table 2. Separate 5 × 2 ANOVAs for each exam were computed. On the end-of-the-semester exam, students' performance on quizzed items (79%) was significantly greater than on nonquizzed items (72%), $F(1, 91) = 45.43$, $\eta_p^2 = .33$ (note that nonquizzed items were tested on prior unit exams). Performance significantly varied across units, $F(4, 364) = 27.35$, $\eta_p^2 = .23$, but there was no interaction between units and quiz condition, $F < 1$. On the end-of-the-year exam, students' performance on quizzed items (68%) remained significantly greater than on nonquizzed items (62%), $F(1, 91) = 4.50$, $\eta_p^2 = .05$. Students' performance at the end of the year also varied across units, $F(4, 364) = 14.95$, $\eta_p^2 = .14$, but again, the interaction between units and quiz conditions was not significant, $F(4, 364) = 1.09$, $p > .05$. Thus, the positive benefits of quizzing were demonstrated over a retention interval that extended for up to 9 months (in the case of the end-of-the-year exam).

Discussion

This experiment and those experiments conducted in social studies classes (Roediger, Agarwal, McDaniel, et al., 2010) are the first to show the effectiveness of low-stakes quizzing in promoting retention of course content on summative assessments used in actual classrooms. Such a finding represents a significant extension over existing experimental work focusing on testing effects, especially the limited research on the testing effect with middle school students. In previous reports of the testing effect with middle school students (and typically college students as well; see Roediger & Karpicke, 2006, for a review), paradigms were used in which either the target material was minimally exposed (presented for students to read once) and was not part of the class curriculum (Spitzer, 1939) or the course material was included in the experiment but the experiment was conducted after students completed their exams and standardized assessments (Carpenter et al., 2009). These parameters are not reflective of typical middle school instructional situations in which the core target (course) content is emphasized by the teacher in her class lectures and learning activities, reinforced in the textbook reading assignments, and tested in summative assessments that count toward course grades. In these situations, students are presumably motivated to study the material (both quizzed and nonquizzed) and perform well on the assessment. We found that even under these favorable conditions for learning target material, students' retention of the material was improved by low-stakes quizzes. Further, the findings indicate that the beneficial quizzing effect lasts at least 9 months, a retention interval substantially longer than previously examined, except for Carpenter et al. (2009; in which quizzing was conducted after the course content had been completed, rather than throughout the course while the material was being learned).

Moreover, from several perspectives, the performance gains on the science content associated with quizzing were impressive and have potentially great practical significance. Quizzing increased students' performance on unit exams from baseline levels of 79% correct (performance when target content was nonquizzed) to levels of more than 90%. The science teacher indicated that the baseline level of performance observed in this study is typical for her eighth-grade science classes, so the baseline is not artificially low. Translated into grades, the quiz-related performance gain represented a change from a C+ grade to an A− grade on the typical grading distribution at that school. Translated into the proportion gains of the unlearned material, quizzing promoted learning of 65% of the material that would otherwise have been answered incorrectly. Essentially the quizzing effects were evidenced for the material that was normatively most difficult to learn—the 21% of the items not correct at baseline.

A second impressive feature of the quizzing effect was that it persisted to the end of semester exam and to the end of the school year. Though the gains were not as substantial as for the unit test, they may be an underestimate of the benefits of testing over long retention intervals. Note that the students' baseline performance for the semester and end-of-the-year exams was based on items on which they had been tested on the unit exams. Thus, students' performance on these baseline items for the long-term retention tests would presumably have benefitted from previous testing (albeit without the feedback provided with quizzes). Another factor may also have been at play, however, in the observed persistence of the quizzing effect. Retrieval on the unit exam of previously quizzed material may have contributed to the long-term retention of the quizzed material. The idea here is that without the additional retrieval on the unit exam, the long-term retention associated with quizzing would be diminished or even eliminated.

We explored this possibility by examining the effect of quizzing on the semester exam separately for items correctly answered and for items not correctly answered on the unit tests. For items correctly answered on the unit tests, semester-exam performance was higher for quizzed items compared with nonquizzed items (.81 vs. .76, respectively), $t(77) = 3.34$, $d = .65$. For incorrect items on the unit exam, students' semester-exam performance did not significantly differ on quizzed and nonquizzed items (.45 vs. .47, respectively; $t < 1$). Thus, the long-term quizzing effect may hinge on additional retrieval on unit exams, though this conclusion is tentative because the needed baseline condition (semester exam performance without preceding unit exams) could not be included in the present classroom environment.

It is also worth emphasizing that the quizzes were low stakes, as scores on the postlesson and review quizzes only counted for 10% of the students' grades. One practical benefit is that such quizzes do not necessarily increase the amount of evaluative testing incorporated into the classroom. This teacher simply chose to use the scores from the quizzes as a means to lessen the reliance on exams for grades. Moreover, on an end-of-the-year survey (completed by 85 of the 92 students included in the previous analyses), students reported that taking the quizzes, especially with clickers, reduced anxiety before a unit exam (64% of respondents) and increased learning (89% of respondents). In line with these survey data, informal observation of the classrooms indicated that students expressed disappointment on the days when the clicker quizzes were not included in the class.

One straightforward interpretation of the current quizzing effect is that the benefits were a direct consequence of testing per se. Another possibility is that the low-stakes quizzes stimulated students to engage in more outside-class work or study, as might be the case for graded quizzes (see Leeming, 2002). However, according to informal observations from the teacher and student surveys, students in the current experiment studied the same amount (69% of respondents) or less (27% of respondents) for this science class in comparison to their other classes. We will defer further consideration of the theoretical underpinnings of these effects until the General Discussion.

Experiments 2a and 2b

Having demonstrated that giving quizzes three times substantially increased exam performance on content in a middle school science class, we were next interested in a more fine-grained evaluation of how the placement of a quiz and the number of repetitions of a quiz would affect the magnitude of this testing effect. The effect of placement of quizzes with respect to the classroom lecture is particularly interesting from both theoretical and practical perspectives. Consider first the potential effects of a quiz that is administered prior to the class lecture on the content (but subsequent to assigned reading). Here students would not be expected to perform exceptionally well on the quiz; however, the quiz would function as prequestions for target materials. In laboratory research, mnemonic benefits of prequestions have been documented (e.g., see Pressley, Tanenbaum, McDaniel, & Wood, 1990; Richland, Kornell, & Kao, 2009). These benefits may in part arise because the quiz items orient the learner toward important content in the subsequently presented material (Mayer, 2003). However, answering the questions produces more learning for students than being exposed to the same prequestions (without having to provide an answer), and this is the case even when prequestions are answered incorrectly (Pressley et al., 1990; Richland et al., 2009). More generally, basic memory research suggests that failing to answer a test question can potentiate learning of the correct answer when it is later provided (Izawa, 1970; Karpicke, 2009; Kornell, Hays, & Bjork, 2009). Thus, the implication here is that though prelecture quizzes may not produce highly accurate performance, they will help students to learn because even the unsuccessful attempts to answer questions potentiate learning of the target material (relative to no prelecture quiz). Although this practical implication has been promoted by researchers (e.g., Kornell et al., 2009), it has not been experimentally evaluated in an educational setting with actual course content.

Another standard placement of a quiz is after a class lecture. Postlecture quizzes can be administered immediately after the class lecture or as a review prior to the summative assessment. We examined both possibilities in the present study. Theoretically, these quiz placements would be expected to promote learning by affording students the opportunity for effective retrieval of target content (Carpenter & DeLosh, 2006; McDaniel & Masson, 1985). Further, for a number of reasons, the review quiz administered just before the summative assessment might be expected to produce the most robust learning benefits. The review quiz would ensure spaced learning of the target material (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006); the target material would have been recently exposed relative to the summative assessment, and the difficulty of retrieval would presumably be greater for the review than immediate postlesson quiz (Bjork, 1994; McDaniel & Masson, 1985).

To investigate these possibilities, we varied placement of certain items in Experiments 2a and 2b across the three types of quizzes (prelesson, postlesson, and review) used in Experiment 1. All students were administered three quizzes (as in Experiment 1), but the quizzed target content included on these quizzes was manipulated such that some content was presented only on the prelesson quiz, some content was presented only on the immediate postlesson quiz, and some content was presented only on the review quiz (the quiz just prior to the summative assessment). This design allowed us to identify whether a single quiz exposure would enhance summative test performance relative to content on which the students had not been quizzed and whether one particular placement of that exposure (prelesson, postlesson, or review) was superior to the others in promotion of test performance. In addition, our design controls for indirect effects of taking quizzes at the three times, because all students received all three quizzes (with content varying on each quiz to instantiate the conditions).

Our second objective in the following two experiments was to explore whether particular combinations of quiz placements would increase the magnitude of the testing effect relative to that obtained with an effective single quiz (assuming that one or more of the single-quiz placements proved to have positive effects). Accordingly, we also implemented the possible two-quiz combinations and the three-quiz combination (see Table 3 for the complete design). In laboratory work, repeated retrieval practice (multiple quizzes) increases the learning benefits from quizzing (Karpicke & Roediger, 2007, 2008). However, as noted earlier, in these laboratory situations, the initial presentation of the target information is typically minimal and less enriched than in a classroom setting

Table 3
Within-Subjects Quiz Conditions Used in Experiments 2a and 2b

Quiz condition	Initial quiz			Unit exam
	$Time_1$	$Time_2$	$Time_3$	
Nonquizzed				X
Prelesson-only	X			X
Postlesson-only		X		X
Review-only			X	X
Prelesson–postlesson	X	X		X
Prelesson–review	X		X	X
Postlesson–review		X	X	X
Prelesson–postlesson–review	X	X	X	X

Note. $Time_1$ refers to a quiz that occurred immediately before the teacher's lesson; $Time_2$ refers to a quiz that occurred immediately after the teacher's lesson, and $Time_3$ refers to a quiz that occurred 24 hr before the unit exam.

(and specifically in the classrooms that participated in the current study). It may be that for the classroom setting, particularly the middle school classes investigated here, repeated quizzing is unnecessary. Nevertheless, in experiments conducted in a college web-based course, McDaniel, Wildman, and Anderson (2010) found that multiple-choice quizzes on core content produced significant benefits on summative-assessment performance primarily when the quizzes were repeated several times. In light of this preliminary finding and the related laboratory research, an alternative expectation is that learning levels will increase as the number of quizzes increases.

We conducted Experiment 2a with the same set of students as Experiment 1. In Experiment 2b, we replicated Experiment 2a with a different set of students and different science content. In addition, Experiment 2b included end-of-the-semester and end-of-year exams. We discuss both sets of results following presentation of each experiment.

Experiment 2a

Method

Participants. The same eighth-grade science students from Experiment 1 ($N = 139$) participated in this experiment.

Materials and design. There were three initial quiz phases: prelesson (before the teacher's lesson but after the students had read an assigned chapter from the textbook), postlesson (after the teacher's lesson about a chapter), and review (24 hr before the unit exam). A 2 (prelesson: quizzed, nonquizzed) × 2 (postlesson: quizzed, nonquizzed) × 2 (review: quizzed, nonquizzed) within-subjects design was used, which created eight quiz conditions. As shown in Table 3, the eight quiz conditions were as follows: (a) a nonquizzed control condition in which items were not presented on initial quizzes but were tested on the unit exam; (b) a prelesson-only condition in which items were presented only on the prelesson quiz; (c) a postlesson-only condition in which items were presented only on the postlesson quiz; (d) a review-only condition in which items were presented only on the review quiz; (e) a prelesson–postlesson condition in which items were presented on both the prelesson and postlesson quizzes; (f) a prelesson-review condition in which items were presented on both the prelesson and review quizzes; (g) a postlesson-review condition in which items were presented on both the postlesson and review quizzes; and finally (h) a prelesson–postlesson–review condition in which items were presented on all three initial quizzes. Note that the quiz conditions reflect the particular combination of quizzes on which any particular quiz item appeared. That is, students received all three quizzes (prelesson, postlesson and review), and each quiz contained a mixture of the quiz items across the quiz conditions.

Forty-eight multiple-choice questions with four alternatives were created from an astronomy unit, with six items per quiz condition. The classroom teacher approved all multiple-choice questions, answers, and lures created by the experimenter. Items were randomly assigned to the eight quiz conditions, and adjustments were made to ensure that every item appeared in every condition no more than once across the six class sections. Thus, each of the six class sections ($M = 24$ students) received a different selection of items in each of the quiz conditions.

The astronomy unit was divided into four chapters, and classroom lessons about chapters varied in length from 3 to 8 days. The experimenter administered a total of four prelesson and four postlesson quizzes (i.e., one before and after each chapter lesson), and the number of items on each quiz varied with respect to the chapter. When collapsed over conditions and chapters, the total number of items on the prelesson quizzes was 24, and the total number of items on the postlesson quizzes was 24. The review quiz, which included material from all four chapters, included 24 items and occurred the day before the unit exam. (Note that the items that appeared on the prelesson, postlesson, and review quizzes did not completely overlap, reflecting the manipulation of quiz–item frequency.) The unit exam consisted of all 48 items, which occurred 31 days after the first prelesson quiz. As in Experiment 1, the teacher included other questions types on the exam as well (fill-in-the-blank questions on the planets and identification of solar or lunar eclipses from images).

Procedure. The same procedure for initial quizzes and the unit exam in Experiment 1 was used in this experiment. End-of-the-semester and end-of-the-year exams were not administered.

Results

Eighteen students who qualified for special education or gifted programs were excluded from our analyses. Furthermore, 56 students who were not present for all initial quizzes and the unit exam were also excluded to ensure the integrity of our quizzing schedule. Therefore, data from 65 students are reported below. However, the general pattern of results remained the same when the 56 absent students were included (see Appendix B, Table B1). In accordance with our primary interests, data have been collapsed over the four chapters of the astronomy unit. All results were significant at an alpha level of .05 unless otherwise noted.

Initial quiz performance. Initial quiz performance as a function of quiz condition is shown in Table 4. As expected, overall performance increased from the prelesson (56%) to the postlesson (77%) and review quizzes (74%). A repeated-measures ANOVA with quiz placement (prelesson, postlesson, and review) as the independent variable revealed a significant effect of quiz placement on initial quiz performance, $F(2, 128) = 96.33$, $\eta_p^2 = .60$. Overall postlesson and review quiz performance were both signif-

Table 4
Students' Average Initial Quiz Performance as a Function of Quiz Condition in Experiment 2a

	Initial quiz performance					
	Prelesson		Postlesson		Review	
Quiz condition	*M*	*SD*	*M*	*SD*	*M*	*SD*
Nonquizzed						
Prelesson-only	.51	.22				
Postlesson-only			.77	.23		
Review-only					.69	.22
Prelesson–postlesson	.57	.19	.76	.20		
Prelesson–review	.55	.21			.73	.19
Postlesson–review			.75	.20	.73	.17
Prelesson–postlesson–review	.62	.21	.81	.19	.82	.18
Average	.56	.13	.77	.14	.74	.12

icantly greater than prelesson quiz performance, $t(64) = 12.85$, $d = 1.55$, and $t(64) = 11.10$, $d = 1.46$, respectively. Performance across the postlesson and review quiz placements did not differ significantly, $t(64) = 1.62$, $p > .05$.

Note that the analysis collapsed postlesson (and review) quiz performances across items that had been previously quizzed and items not previously quizzed. Accordingly, to distinguish the potential contributions of prelesson quizzing versus simply the teacher's lesson, we next examined postlesson quiz performance as a function of whether the content was present on the prelesson quiz (postlesson-only vs. prelesson–postlesson). This analysis indicated no significant differences on postlesson performance as a function of whether the information had appeared on the prelesson quiz (78% when quizzed prelesson vs. 76% when not quizzed prelesson), $t(64) = 1.14$, $d = 0.15$, $p > .05$. Similarly, on the review quiz, performance for prelesson-review items (73%) did not significantly differ from performance on review-only items (69%), $t(64) = 1.48$, $p > .05$. These results indicate that students' improvements from prelesson to postlesson and review quizzes were largely a consequence of students' learning from the teacher's lesson.

Unit exam performance. Unit exam performance as a function of quiz condition is shown in Table 5. Our first set of analyses related to the issue of how the placement of a quiz influences its impact on students' summative test performance. To isolate the effects of quiz placement per se, we analyzed unit exam performance as a function of the single quiz conditions and the nonquizzed items. A one-way within-subjects ANOVA of the nonquizzed, prelesson-only, postlesson-only, and review-only quiz conditions revealed a significant effect, $F(3, 192) = 21.09$, $\eta_p^2 = .25$, indicating that the placement of a single quiz significantly affected final-unit exam performance. Planned *t* tests comparing each type of quizzed item to the nonquizzed items demonstrated significant testing effects for postlesson-only quizzed items (77% relative to 64% for nonquizzed items), $t(64) = 3.89$, $d = 0.66$, and the review-only quizzed items (86% vs. 64%), $t(64) = 7.22$, $d = 1.13$. The prelesson-only quizzed items did not show a significant advantage (69%) relative to nonquizzed items (64%), $t(64) = 1.71$, $d = 0.22$, $p = .09$. In addition, performance on the review-only quizzed items (86%) was significantly greater than performance on the postlesson-only quizzed items (77%), $t(64) = 2.91$, $d = 0.47$.

Our second set of analyses focused on whether some combinations of quizzing would be more potent than a single-quiz presentation. Although items appearing only on a prelesson quiz were found to provide little benefit on the final examination (in comparison to the nonquizzed condition), it may be that a prelesson quiz combined with another quiz would aid retention over and above that obtained with postlesson and review quizzes alone. This possibility was not supported, however. Planned *t* tests indicated that the prelesson–postlesson quizzing condition (78%) did not increase students' performance on the unit exam relative to postlesson-only quizzing (77%), $t < 1$, $p > .05$, and the prelesson–review quizzing condition (85%) did not increase students' performance on the unit exam over the review-only quizzing (86%), $t < 1$, $p > .05$. Thus, in the present learning context, a single quiz administered before the teacher's lesson (i.e., the prelesson quiz) did not enhance learning, either when presented alone or in combination with quizzes that occurred after the teacher's lesson (i.e., the postlesson and review quizzes).

We also investigated whether the potency of the single most effective quiz (review quizzing) was augmented with additional prior quizzing. Planned *t* tests showed that combining the postlesson and review quizzing (87%) was not significantly more effective than review-only quizzing (86%), $t < 1$, $p > .05$; even when prelesson, postlesson, and review quizzing were administered (89%), there was no significant improvement on the examination relative to review-only quizzing (86%), $t(64) = 1.39$, $p > .05$. Thus, additional quizzing did not significantly augment the benefits of a review quiz on students' performance on the unit examination. Of course, the performance levels in these last analyses are near ceiling, possibly reducing the sensitivity to detect differences.

Experiment 2b

In this experiment, we sought to replicate Experiment 2a and extend those results to delayed retention exams administered at 3 months (end of semester) and 8 months (end of year) after initial learning.

Method

Participants. Eighth-grade science students ($N = 148$) from the same middle school as in prior experiments partici-

Table 5
Students' Average Unit Exam Performance as a Function of Quiz Condition in Experiment 2a

	Unit exam performance	
Quiz condition	*M*	*SD*
Nonquizzed	.64	.21
Prelesson-only	.69	.22
Postlesson-only	.77	.19
Review-only	.86	.17
Prelesson–postlesson	.78	.20
Prelesson–review	.85	.17
Postlesson–review	.87	.14
Prelesson–postlesson–review	.89	.13

pated in this experiment. Parents were informed of the study, and written assent from each participant was obtained in accordance with the Washington University Human Research Protection Office. Participants in Experiment 2b did not participate in the prior experiments.

Materials and design. The design used in Experiment 2a was the same as that used in Experiment 2b, which included two units from the eighth-grade science curriculum: genetics, and chemistry. Fifty-six multiple-choice genetics questions and 48 multiple-choice chemistry questions, all with four alternatives, were created by the experimenter and approved by the classroom teacher. At the outset, we planned to collapse data across the two units, yielding a total of 104 multiple-choice items, or 13 items in each of the eight quiz conditions.

Both units were divided into four chapters each, and classroom lessons about chapters varied in length between 1 and 4 days. The experimenter administered a total of eight prelesson and eight postlesson quizzes (i.e., one before and after each chapter lesson, respectively), and the number of items on each quiz varied with respect to the chapter. When collapsed over conditions, chapters, and units, the total number of items on the prelesson quizzes was 52, and the total number of items on the postlesson quizzes was 52. Review quizzes occurred the day before the unit exam, also with a total of 52 items. Retention was measured on unit exams, including all 104 items, which were administered an average of 31 days after the first prelesson quiz for both units. As in the previous experiments, the unit exams also consisted of nonmultiple choice questions (e.g., students were asked to label parts of the atom and fill out information about various elements using the periodic table) and a few multiple choice questions (see description to Experiment 1 method) that were not included in the experimental analyses.

Two additional delayed-retention exams were used: a 3-month delayed semester exam and an 8-month delayed year exam. For the semester exam, 32 items were randomly selected from each unit (as in Experiment 1, the semester exam also included items from units not targeted for this experiment). Therefore, 64 items were selected over both units, with eight items per condition. For the delayed year exam, 10 items from the genetics unit were randomly selected, five items from the nonquizzed condition, and five items from the prelesson–postlesson-review condition. Items from the chemistry unit were not included on the delayed year exam due to classroom time constraints. On both of the delayed exams, items were presented in random order within unit, and units were presented in the chronological order in which they were taught.

Procedure. The same procedure used for initial quizzes and the unit exam in Experiment 2a was used in Experiment 2b. Regarding the delayed semester exam, students were notified of the date of the exam approximately 1 month before it was administered, as performance was recorded for grading purposes. Items on the delayed semester exam were presented in a test booklet, and students recorded their answers on a Scantron form (Scantron Corp., Eagan, MN). Regarding the delayed year exam, students were not informed of the exam until it was administered. Items on the delayed year exam were presented on the same clicker system used during initial quizzes.

Results

Twenty-five students who qualified for special education or gifted programs were excluded from our analyses. Furthermore, 69 students who were not present for all initial quizzes, unit exams, and delayed exams were also excluded to ensure the integrity of our quizzing schedule. Therefore, data from 54 students are reported below. However, despite these exclusions, the general pattern of results remained the same when the 69 absent students were included (see Appendix B, Table B2). In accordance with our primary interests, data were collapsed over the genetics and chemistry units. All results were significant at an alpha level of .05 unless otherwise noted.

Initial quiz performance. Initial quiz performance as a function of quiz condition is shown in Table 6. As expected, there was a significant main effect of type of quiz (prelesson, postlesson, and review) on initial quiz performance, $F(2, 106) = 285.64$, $\eta_p^2 = .84$. Overall postlesson (84%) and review quiz (84%) performances were not different ($t < 1$), and both were significantly greater than prelesson performance (62%), $t(53) = 19.37$, $d = 2.27$ and $t(54) = 20.12$, $d = 2.44$, respectively. These results replicate Experiment 2a.

As in Experiment 2a, to distinguish the potential contributions of prelesson quizzing versus the teacher's lesson on students' later quiz performance, we also examined postlesson quiz performance as a function of whether the content was present on the prelesson quiz (postlesson-only vs. prelesson–postlesson). Unlike Experiment 2a, prelesson–postlesson performance on the postlesson quiz (84%) was significantly greater than postlesson-only performance (79%), $t(53) = 2.06$, $d = .35$. However, on the review quiz, performance for prelesson–review items (81%) did not significantly differ from performance on review-only items (77%), $t(53) = 1.71$, $p > .05$. Thus, while prelesson quizzing may produce a benefit for postlesson performance, this benefit did not persist to the review quiz.

Unit exam performance. Unit exam performance as a function of quiz condition is shown in Table 7. Following Experiment 2a, we first isolated the influence of the placement of a single quiz. A one-way ANOVA between the nonquizzed items (83%), prelesson-only items (86%), postlesson-only items (89%), and review-only items (92%) revealed a significant effect, $F(3, 159) =$

Table 6
Students' Average Initial Quiz Performance as a Function of Quiz Condition in Experiment 2b

	Initial quiz performance					
	Prelesson		Postlesson		Review	
Quiz condition	*M*	*SD*	*M*	*SD*	*M*	*SD*
Nonquizzed						
Prelesson-only	.60	.15				
Postlesson-only			.79	.13		
Review-only					.77	.14
Prelesson–postlesson	.61	.16	.84	.14		
Prelesson–review	.64	.14			.81	.14
Postlesson–review			.81	.15	.87	.11
Prelesson–postlesson–review	.62	.12	.90	.08	.93	.07
Average	.62	.10	.84	.09	.84	.08

Table 7
Students' Average Unit Exam Performance as a Function of Quiz Condition in Experiment 2b

Quiz condition	Unit exam performance	
	M	*SD*
Nonquizzed	.83	.14
Prelesson-only	.86	.11
Postlesson-only	.89	.10
Review-only	.92	.08
Prelesson–postlesson	.89	.11
Prelesson–review	.94	.08
Postlesson–review	.94	.07
Prelesson–postlesson–review	.96	.06

10.42, $\eta_p^2 = .16$, indicating the utility of a single quiz on final unit exam performance. Planned *t* tests demonstrated significant testing effects when we compared the postlesson-only and nonquizzed conditions, $t(53) = 3.94$, $d = .51$, and the review-only and nonquizzed conditions, $t(53) = 4.27$, $d = .76$. The prelesson-only quiz did not significantly improve unit exam performance relative to nonquizzed items, $t(53) = 1.35$, $p > .05$, similar to Experiment 2a.

In a second set of analyses, we focused on whether some combinations of quizzing would be more potent than a single quiz presentation. Planned *t* tests comparing the prelesson–postlesson (89%) with postlesson-only (89%) conditions and the prelesson–review (94%) with review-only (92%) conditions revealed no significant advantage of combining a prelesson quiz with any other quiz (relative to giving the other quiz alone), ts < 1.88, ps $> .05$. Consistent with Experiment 2a, a single quiz given before the teacher's lesson was not effective in enhancing students' performance on the unit exam, and this type of initial quiz did not provide any additional benefit to performance on quizzes that occurred after the teacher's lesson.

The next set of planned *t* tests focused on the review-only quiz, which resulted in the greatest unit exam performance relative to the other single-quiz exposure conditions (and to nonquizzed items). In contrast to Experiment 2a, the benefit from review quizzes was increased with additional quizzes. The postlesson–review (94%) and the prelesson–postlesson–review (96%) conditions produced significantly higher exam performance than the review-only (92%) condition, ts > 2.43, ps $< .05$. Thus, additional quizzes may augment the benefits of a review quiz on unit exam performance.

Delayed exam performance. End-of-semester and end-of-the-year exam performances as a function of quiz condition are shown in Table 8. Recall that items on the end-of-semester exam were selected from both the genetics and chemistry units, and students were tested once on all items (including the nonquizzed items) on the unit exam. Items on the end-of-the-year exam were selected only from the genetics unit (and only from the prelesson–postlesson–review and nonquizzed conditions, due to time constraints), and students were tested on these items once on the unit exam but were not tested on them on the semester exam.

For the end-of-semester exam, a one-way ANOVA between the nonquizzed (76%), prelesson-only (81%), postlesson-only (81%), and review-only (86%) quiz conditions demonstrated a significant effect, $F(3, 159) = 6.52$, $\eta_p^2 = .11$, indicating the utility of a single quiz (followed by a unit exam) on students' performance on the 3-month delayed semester exam. We conducted planned *t* tests that demonstrated significant testing effects when the postlesson-only and nonquizzed conditions were compared, $t(53) = 2.20$, $d = .32$, and the review-only and nonquizzed conditions were compared, $t(53) = 4.57$, $d = .61$. The prelesson-only and nonquizzed comparison was only marginally significant, $t(53) = 1.88$, $d = .29$, $p = .07$. Thus, even after a 3-month delay, robust testing effects were obtained following the postlesson and review single quiz conditions.

The review-only condition resulted in significantly greater performance on the end-of-semester exam than the prelesson-only and postlesson-only conditions, ts > 2.11, ps $< .05$, further demonstrating the enduring effect of review quizzing the day before the unit exam. There were no significant differences among the conditions that included a review quiz.

Finally, regarding the 8-month delayed year exam, which consisted of only genetics items, a substantial testing effect was revealed such that the prelesson–postlesson–review condition (82%) resulted in significantly greater performance than the nonquizzed condition (69%), $t(53) = 3.87$, $d = .61$. This result demonstrates the robust effect of quizzing after a very long delay (at least when quizzes are followed by unit exams). Although we could not include items from all eight quiz conditions on the delayed year exam, we speculate that the review-only condition would have provided a similar testing benefit on the delayed year exam on the basis of consistent results across Experiments 1 and 2a.

Discussion of Experiments 2a and 2b

Overall, the results from Experiments 2a and 2b converge on a number of conclusions and provide important insights into an effective schedule for quizzing. First, significant testing effects were found with two single-quiz conditions. Both the postlesson and review quizzes alone resulted in greater unit and semester exam performance than the nonquizzed condition. Accordingly, multiple low-stakes quizzes are not necessary for producing a learning or retention benefit on classroom exams. Middle school teachers with limited class time for administering quizzes could quiz students on core material once (if quizzed after lesson cov-

Table 8
Students' Average Delayed Exam Performance as a Function of Quiz Condition in Experiment 2b

Quiz condition	Delayed examination performance			
	End-of-the-semester		End-of-the-year	
	M	*SD*	*M*	*SD*
Nonquizzed	.76	.18	.69	.03
Prelesson-only	.81	.15		
Postlesson-only	.81	.15		
Review-only	.86	.14		
Prelesson–postlesson	.83	.02		
Prelesson–review	.81	.02		
Postlesson–review	.83	.02		
Prelesson–postlesson–review	.85	.02	.82	.02

erage) and still expect to enhance their students' learning and retention of the content.

Second, of the three single-quiz placements investigated here (prelesson, postlesson, and review), the review quiz consistently resulted in the greatest unit and semester exam benefits (relative to nonquizzed content). Of course, the effect on unit exams could have been due to the fact that the review quiz occurred shortly before the unit test. However, this was not the case for the end-of-the-semester exam (though as noted in discussion to Experiment 1, retrieval on the unit exam may have conferred additional benefit). Also, additional quizzes tended not to increase the review quiz benefit. Only in Experiment 2b did additional quizzes augment the review quiz effect on the unit exams, and this pattern did not last over the 3-month retention interval on the end-of-the-semester exam. Thus, a single low-stakes review quiz produces potent testing effects (at least for middle school science content), effects that are augmented only slightly, if at all, by additional quizzes administered prior to the review quiz.

Theoretically, the advantages of the review quiz may involve several factors. First, the greater delay between the initial lesson and the review quiz (relative to a postlesson quiz, which occurred immediately after the lesson) creates a more challenging retrieval condition, which may enhance the mnemonic benefits of retrieval (Bjork, 1994; McDaniel & Masson, 1985). We note the potential retrieval effects associated with the delay, rather than potential effects of the feedback, because performance on both the postlesson and review quizzes were high and not different (indicating generally successful retrieval of the content). Second, the review quiz spaces the material from the lesson more so than the other quiz placements did, and such spacing could confer an advantage (Cepeda et al., 2006). Third, as noted earlier, the review-quiz advantage could be related to the shorter retention interval between the review quiz and the unit exam (relative to the postlesson quizzes administered some days before the exam), which would result presumably in students forgetting less of the quizzed content. Clearly, all factors could be involved, and this finding points to an issue to be examined further in controlled laboratory studies.

However, it is worth emphasizing that the review quiz conferred advantages for performance (relative to no-quizzing, prelesson quizzing, and postlesson quizzing) that extended well beyond the unit exam to an end-of-the-semester exam (perhaps assisted in part by retrieval on the unit exam). Students were aware of this end-of-the-semester exam, which included items from the unit exam, and thus items from all conditions were on equal footing in terms of having been identified as possible test items to review for end-of-semester study (cf. Carpenter et al., 2009). This may indicate that either the retrieval dynamics associated with a review quiz or more spacing of content is involved in its effectiveness. Regardless, Experiment 2b suggests that a single review quiz administered a day before a final unit exam provides the largest benefit for students' long-term retention relative to the other quiz placements investigated here and relative to no quizzing. It remains to extend this effect to other quiz formats (e.g., short answer), other content, and students at different educational levels.

A third important finding is that quizzes given before the teacher's lesson (the prelesson quizzing condition) did little to enhance final retention, even when administered in conjunction with other quizzes. This result is inconsistent with practitioners' impressions that a pretest can facilitate learning by orienting students toward upcoming material and perhaps also activate any relevant prior knowledge that may help scaffold the lesson contents. These results also do not support laboratory experimental findings that have reported benefits of prequestions on learning (Pressley et al., 1990; Richland et al., 2009). Numerous differences exist between the current educational context and laboratory studies that could explain the apparent discrepancy in findings. For instance, possibly the current quiz questions were not gauged appropriately to provide an assumed orienting or scaffolding function; perhaps performance on the current summative assessment (multiple-choice questions) did not require the degree of organizational encoding that laboratory summative assessments might (e.g., requiring recall; Pressley et al., 1990); or possibly the current content was more novel to the students than content used in laboratory research.

In addition, it does not seem to be the case that unintended chance differences in item difficulty totally accounted for the limited benefits of prelesson quizzes. This possibility arose in Experiment 2a especially, with few items in each quiz condition; here, prelesson quiz performance was lowest on the prelesson-only items (.11 lower than performance on the prelesson quiz for the items appearing on the prelesson–postlesson–review quiz), and in turn unit exam performance was lowest for this condition. However, the reverse occurred for the prelesson–postlesson and prelesson–review conditions, with lower prelesson quiz performance associated with higher unit exam performance. Moreover, in Experiment 2b, students' performance on prelesson-only items was nearly equivalent to their performance on items on the prelesson quiz in the other quiz conditions, yet the prelesson-only quiz still did not produce unit exam levels comparable to those of the other quiz conditions. Accordingly, though the random assignment of items to particular quiz conditions produced some variation in item difficulty across quiz conditions (as would be expected), overall this did not appear to be systematically associated with performance levels on unit exams.

As researchers, we were guests in a teacher's classroom, and accordingly we were unable to control grading policies. As mentioned, the postlesson and review quizzes, but not the prelesson quizzes, often counted for a small percentage of the students' overall grade. This situation raises the possibility that motivational effects attenuated the potential benefit of the prelesson quizzes. However, this factor seems unlikely to explain our differences for several reasons. First, the importance of the prelesson quizzes was always stressed, and the scores were displayed after the quiz. Students were well aware that these questions were likely going to be included on their later quizzes and tests, so they would certainly have reason to pay attention to the feedback, the content in the readings, and the teacher's lectures. Also, although quizzing only on the prelesson quiz did not enhance performance on the unit exam, there is some evidence that performance on the subsequent (postlesson) quiz benefitted slightly. Clearly, the current study was not designed to identify the critical factors that promote benefits of prelesson quizzing. The results do suggest, however, that the benefits for prelesson quizzing may not be especially widespread for authentic educational situations with the kinds of course content and summative examinations used here.

Finally, a robust testing effect was demonstrated after an 8-month delay on the end-of-year exam (Experiment 2b). This effect was evaluated only for content that was quizzed three times

and is thus consistent with that found in Experiment 1. This finding converges with those reported in a laboratory experiment in which testing effects were found for educational content (art history) after a 2-month delay (Butler & Roediger, 2007) and in a classroom experiment in which testing effects were found for eighth-grade history facts after a 9-month delay (Carpenter et al., 2009). The present finding extends these effects to an authentic educational setting (unlike Butler & Roediger), and one in which low-stakes quizzes were administered during the presentation of the unit material (unlike Carpenter et al. in which the quiz was administered as an experiment after the summative assessments for those facts had been completed).

In considering these long-term effects of quizzing, as discussed in Experiment 1, we note that students also were tested on the nonquizzed items on the unit exams, and thus they also may have benefited from being tested on this exam (though the present study was not designed to gauge a benefit of the unit test per se). Quizzes may have promoted long-term retention because they provided more frequent retrieval practice (including practice accruing from taking the unit exam) than did the unit exam alone. However, the quizzes were also accompanied by corrective feedback, whereas the unit tests were not. Thus, the long-term effects of quizzing could be at least in part related to benefits of feedback. More fine-grained experimental work will be needed to directly inform these possibilities.

General Discussion

Although the positive benefits of testing (quizzing) on retention and subsequent test performance have been documented in laboratory experiments (for a review, see Roediger & Karpicke, 2006) and with educationally relevant materials (e.g., Butler & Roediger, 2007; Carpenter et al., 2009; Glover, 1989), there is a paucity of controlled classroom experiments on the testing effects on the curricular material that students are required to learn for their course assessments (and by extension for high-stakes achievement tests). As discussed previously, important differences exist between the classroom and laboratory contexts, including the repeated emphasis on to-be-learned content and presumably greater motivational dynamics for learning in the classroom context. Because these differences could possibly reduce or eliminate the testing effect, uncertainty has remained regarding the potency of testing effects in authentic classroom settings. The three experiments reported herein provide strong evidence that for middle school science classes, low-stakes quizzing with feedback can be extremely effective in achieving one educational objective—that of increasing performance on summative assessments (for learning and retention) of target content.

When three quizzes on the content were spaced across the coverage of a unit, effects were robust, producing between 13% and 25% gains in performance on unit exams across a range of eighth-grade science topics, including genetics, evolution, anatomy, astronomy, and chemistry. Perhaps more telling, the proportion of otherwise unlearned material (i.e., material with incorrect answers for nonquizzed items) that benefited from quizzing was substantial, ranging from .61 (Experiment 2a) to .78 (Experiment 2b, for the three-quizzed conditions). From the standpoint of projected grades, the benefits of quizzing were equally impressive. Performance levels on material that was not quizzed were at the C+ level for the school's grading scale; quizzing generally increased performance levels on the material to an A– level. This is a robust effect, especially considering that the science teacher for this grade level provided conscientious, rich instruction for her students, with the classroom experience involving active learning exercises and demonstrations. The summative exams reflected the content emphasized in the textbook and the classroom activities. Even with these highly favorable pedagogical factors in place, quizzing improved learning and exam performance for the course content relative to no quizzing.

Further, the present quizzing effects might slightly underestimate the true effect because quizzing can augment students' performance on related content (Chan, McDermott, & Roediger, 2006), and we cannot measure that influence in our design. Some (but not a majority) of the exam items were on related content (e.g., in genetics, two questions focused on meaning of *phenotype*, and in astronomy, two questions focused on the related concepts of equinox and solstice), and to the extent that in some cases random assignment distributed one of each into quiz and nonquiz conditions, then there would be a chance for some carryover effects to nonquizzed items. If this happened, the direct effect of quizzing would be underestimated because the control comparison would be too high (due to the spillover effect of quizzing on related items).

The quizzing effects are also impressive from the standpoint of the minimal adjustments required to incorporate them into the classroom. The introduction of the low-stakes quizzing into this science classroom required only minor changes to the normal classroom practices. The teaching approach did not change, the course materials did not change, and the curriculum did not change. The primary changes for the teacher involve constructing the quizzes and setting aside some class time (on the order on minutes) for administering the quizzes. Indeed, it appears that the class time needed to obtain significant benefits of quizzing could be relatively small. Experiments 2a and 2b showed that just one strategically placed quiz on target content (after a lesson or as a review for a unit exam) produced significant increases in unit exam performance. Further, when a single quiz was administered as a review activity a day prior to the unit exam, the benefits were nearly (and statistically) equivalent to benefits from repeating the quiz several times, and the benefits of the single review quiz were long-lasting (up to 3 months; Experiment 2b). Still, with more strategic spacing of repeated quizzes one might expect that repeating quizzes would produce gains above that obtained with a single review quiz; however, the current repeated quizzing scheme was one that had been favored (on the basis of subjective impressions) in a social studies class at the middle school (see Roediger, Agarwal, McDaniel, et al., 2010). Another practical advantage of the present quizzing procedure rests on its low-stakes feature. Because the quizzes were low stakes, feedback could be broadcast to the students in the classroom setting, without requiring additional teacher time for grading. Moreover, the low-stakes nature of the quizzes reduced student anxiety (according to self-reports) and increased student learning. Thus, the cost–benefit ratio for low-stakes quizzing in middle school science classes (and other content domains as well, Roediger, Agarwal, McDaniel, et al., 2010) appears to be highly favorable.

There are a number of possible theoretical explanations for the present effects. Though the present study was not designed to disentangle these explanations, some of the more prominent pos-

sibilities merit mention. Several direct effects of testing were possibly operative. One is an exposure effect; the students received additional exposure to quizzed material relative to nonquizzed material. From the laboratory findings for the testing effect (e.g., Karpicke & Roediger, 2008; McDaniel & Masson, 1985), it seems likely, however, that the present quizzing effects were not entirely a consequence of additional exposure per se. Most pertinent, in the parallel work that we have been conducting in middle school social studies classes, we have found that exposure per se (presenting the target content but not in quiz format) improves performance only on unit exams (not on semester and end-of-the-year exams) and not to the levels obtained with quizzing (Roediger, Agarwal, McDaniel, et al., 2010; see McDaniel et al., 2010, for a similar finding in a university psychology course). Therefore, it is likely that the present effects also reflected at least some benefit accruing to the retrieval processes or retrieval practice required for answering quiz items, benefits that are well established in the basic literature (e.g., Carpenter & DeLosh, 2006; McDaniel & Masson, 1985; Roediger & Karpicke, 2006).

Another possible contributor to the effects could be learning from feedback, which can be potent (e.g., Butler & Roediger, 2008; McDaniel & Fisher, 1991; Pashler et al., 2005), especially when a correct answer is provided for a failed answer (Izawa, 1970; Kornell et al., 2009). This latter situation of learning from feedback for incorrect responses may be most pertinent for the single review quiz condition that showed impressive testing effects (Experiments 2a and 2b). For these single-quiz conditions, incorrect performance ranged from 23% to 33% across experiments; therefore, feedback-driven learning was provided for a fair amount of material (and the material that was most likely to not have been learned for nonquizzed content).

The factors just noted in the previous paragraph may be especially potent because following the preponderance of the laboratory paradigms used to investigate testing effects, the unit exam questions were the same as those used for the quizzes. Thus, retrieval practice and feedback for particular target information were recapitulated perfectly for the unit exam items. Clearly, not all educators will embrace the technique of giving identical items on the quiz and the summative assessment (see e.g., Mayer et al., 2009). Note, though, that the present quizzing method was adopted from that already in use by a social studies teacher at the school. Further, there are many learning contexts in which basic information must be mastered, and retrieval practice on that information prior to the summative exam may be appropriate (see e.g., Larsen, Butler, & Roediger, 2009, in a medical school context; McDaniel et al., 2010, in a neuroscience course context). Still, the educational value of quizzing would be significantly broadened if quizzing improved performance on summative assessments that did not contain the identical quiz questions.

An encouraging development is that initial experimental evidence in the laboratory (Chan et al., 2006; Rohrer, Taylor, & Sholar, 2010) and classroom (McDaniel et al., 2007; McDaniel et al., 2010) contexts, as well as correlational findings in college classes (e.g., Angus & Watson, 2009; Kibble, 2007), has demonstrated positive effects of quizzing, albeit often with reduced magnitudes (and mostly with college students), when the summative assessment questions are not identical to the quiz questions. Butler (2010) has reported four experiments showing that material that has been quizzed leads to greater transfer to new questions than material that was reread (see also McDaniel, Howard, & Einstein, 2009). In related classroom research with clicker technology, college students who answered questions used to assess understanding (formative assessment) and then discussed the rationale for correct answers scored better on class exams than did students not given clicker questions and discussion (Mayer et al., 2009). The effect was observed for exam questions that were similar and dissimilar in content to the clicker questions.

A host of possible indirect effects of testing could contribute to positive quizzing effects. Quizzing might improve metacognition for the course content. Fifty-five percent of the students reported becoming better at assessing what they did and did not understand following our clicker quizzes, and although we did not measure metacognition or study strategies, this increase in metacognition could allow students to more effectively self-direct study activities (e.g., Thomas & McDaniel, 2007). Quizzing could simply stimulate students to keep up better with class assignments (see Leeming, 2002), though given the low-stakes nature of the present quizzing regimen and survey responses this possibility seems less likely in our experiments. The low-stakes aspect of the quizzing reduced test anxiety, and the quizzes could also increase motivation (as the student experiences success on the quizzes). Identifying which, if any, of the possible mechanisms underlie the benefits of quizzing remains for more fine-grained research. What is clear is that at least for middle school science classes, low-stakes quizzing produces significant gains in learning and retention, as assessed by standard classroom summative assessments.

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. doi: 10.1002/acp.1391

Angus, S. D., & Watson, J. (2009). Does regular online testing enhance student learning in the numerical sciences? Robust evidence from a large data set. *British Journal of Educational Technology, 40,* 255–272. doi: 10.1111/j.1467-8535.2008.00916.x

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

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

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

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

Butler, A. C., & Roediger, H. L., III (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36,* 604–616. doi:10.3758/MC.36.3.604

Callender, A. A., & McDaniel, M. A. (2007). The benefits of embedded question adjuncts for low and high structure builders. *Journal of Educational Psychology, 99,* 339–348. doi:10.1037/0022-0663.99.2.339

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative processing 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. doi:10.1002/acp.1507

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132,* 354–380. doi:10.1037/0033-2909.132.3.354

Chan, J. C. K., McDermott, K. B., & Roediger, H. L., III (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General, 135,* 553–571. doi:10.1037/0096-3445.135.4.553

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81,* 392–399. doi:10.1037/0022-0663.81.3.392

Izawa, C. (1970). Optimal potentiating effects and forgetting-prevention effects of tests in paired-associate learning. *Journal of Experimental Psychology, 83,* 340–344. doi:10.1037/h0028541

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

Karpicke, J. D., & Roediger, H. L., III (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language, 57,* 151–162. doi:10.1016/j.jml.2006.09.004

Karpicke, J. D., & Roediger, H. L., III (2008, February 15). The critical importance of retrieval for learning. *Science, 319,* 966–968. doi:10.1126/science.1152408

Kibble, J. (2007). Use of unsupervised online quizzes as formative assessment in a medical physiology course: Effects of incentives on student participation and performance. *Advances in Physiology Education, 31,* 253–260. doi:10.1152/advan.00027.2007

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 989–998. doi:10.1037/a0015729

Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory, 17,* 493–501. doi:10.1080/09658210902832915

Larsen, D. P., Butler, A. C., & Roediger, H. L., III (2009). Repeated testing improves long-term retention relative to repeated study: A randomized controlled trial. *Medical Education, 43,* 1174–1181. doi:10.1111/j.1365-2923.2009.03518.x

Leeming, F. C. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology, 29,* 210–212. doi:10.1207/S15328023TOP2903_06

Mayer, R. E. (2003). Memory and information processes. In W. M. Reynolds & G. E. Miller (Eds.), *Handbook of psychology: Vol. 7. Educational psychology* (pp. 47–57). Hoboken, NJ: Wiley.

Mayer, R. E., Stull, A., DeLeeuw, K., Almeroth, K., Bimber, B., Chun, D., . . . Zhang, H. (2009). Clickers in college classrooms: Fostering learning with questioning methods in large lecture classes. *Contemporary Educational Psychology, 34,* 51–57. doi:10.1016/j.cedpsych.2008.04.002

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19,* 494–513. doi:10.1080/09541440701326154

McDaniel, M. A., & Donnelly, C. M. (1996). Learning with analogy and elaborative interrogation. *Journal of Educational Psychology, 88,* 508–519.

McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology, 16,* 192–201. doi:10.1016/0361-476X(91)90037-L

McDaniel, M. A., Howard, D. C., & Einstein, G. O. (2009). The read-recite-review study strategy: Effective and portable. *Psychological Science, 20,* 516–522. doi:10.1111/j.1467-9280.2009.02325.x

McDaniel, M. A., & Masson, M. E. J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 371–385. doi:10.1037/0278-7393.11.2.371

McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2010). *Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study.* Manuscript submitted for publication.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 3–8. doi:10.1037/0278-7393.31.1.3

Pressley, M., Tanenbaum, R., McDaniel, M. A., & Wood, E. (1990). What happens when university students try to answer prequestions that accompany textbook material? *Contemporary Educational Psychology, 15,* 27–35. doi:10.1016/0361-476X(90)90003-J

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied, 15,* 243–257. doi:10.1037/a0016496

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

Roediger, H. L., III, Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2010). *Test-enhanced learning in the classroom: Long-term improvements from quizzing.* Manuscript submitted for publication.

Roediger, H. L., III, & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1,* 181–210. doi:10.1111/j.1745-6916.2006.00012.x

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 233–239. doi:10.1037/a0017678

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology, 30,* 641–656. doi:10.1037/h0063404

Thomas, A. K., & McDaniel, M. A. (2007). Metacomprehension for educationally relevant materials: Dramatic effects of encoding-retrieval interactions. *Psychonomic Bulletin & Review, 14,* 212–218.

Ward, D. (2007). eInstruction: Classroom Performance System [Computer software]. Denton, TX: eInstruction Corp.

Appendix A

Examples of questions from three units: Experiment 1 (genetics), Experiment 2a (anatomy), and Experiment 2b (chemistry). (F) denotes a factual item; (A) denotes an analytic or inferential item.

Genetics

(F) An organism's physical appearance is its _____.

A. genotype

B. phenotype

C. codominance

D. heterozygous

(F) What are the building blocks of protein?

A. Chromosomes

B. Amino acids

C. DNA

D. RNA

(A) How can genetic counselors predict genetic disorders?

A. By studying karyotypes and pedigree charts

B. By taking pictures of a baby before it is born

C. By exploring new methods of genetic engineering

D. By eliminating codominant alleles in the parents

(A) Why are sex-linked traits more common in males than in females?

A. All alleles on the X chromosome are dominant.

B. All alleles on the Y chromosome are recessive.

C. Males only have one X chromosome, so if they inherit the recessive allele, they will show that particular trait.

D. Any allele on the Y chromosome will be codominant with the matching allele on the X chromosome.

Anatomy (Skeletal, Muscular, and Integumentary [Skin] Systems)

(F) The average human adult has about ____ pounds of skin.

A. 6

B. 12

C. 20

D. 32

(F) Your shoulder and hip are what type of joint?

A. Ball and socket

B. Pivot

C. Hinge

D. Gliding

(A) Weight lifting, sprinting, and doing pushups are all examples of what type of exercise?

A. Aerobic

B. Anaerobic

C. Catabolic

D. Anabolic

(A) Which of the following is controlled by involuntary muscles?

A. Breathing

B. Smiling

C. Talking

D. Walking

Chemistry

(F) A(n) _____ change is when a substance is changed into a different substance with different properties.

A. physical

B. chemical

C. hydrothermal

D. energy

(F) The hardest form of carbon is what?

A. Fullerene

B. Diamond

C. Alloy

D. Graphite

(A) What can you tell from the molecular formula for methane (CH_4)?

A. It contains four carbon atoms

B. It contains one hydrogen atom

C. It contains four hydrogen atoms

D. It forms groups of four molecules

(A) Which of the following is an amorphous solid?

A. Plastic doll

B. Ice castle

C. Copper penny

D. All of the above

(*Appendices continue*)

Appendix B

Table B1
Students' Average Unit Exam Performance as a Function of Quiz Condition in Experiment 2a

	Unit exam performance	
Quiz condition	*M*	*SD*
Nonquizzed	.66	.21
Prelesson-only	.69	.24
Postlesson-only	.75	.21
Review-only	.83	.20
Prelesson–postlesson	.76	.22
Prelesson–review	.82	.20
Postlesson–review	.85	.18
Prelesson–postlesson–review	.86	.17

Note. Included both present and absent students.

Table B2
Students' Average Unit Exam Performance as a Function of Quiz Condition in Experiment 2b

	Unit exam performance	
Quiz condition	*M*	*SD*
Nonquizzed	.82	.15
Prelesson-only	.84	.12
Postlesson-only	.87	.11
Review-only	.90	.10
Prelesson–postlesson	.88	.12
Prelesson–review	.93	.08
Postlesson–review	.93	.08
Prelesson–postlesson–review	.95	.07

Note. Included both present and absent students.

Received April 13, 2010
Revision received September 21, 2010
Accepted October 7, 2010 ■