Applied Cognitive Psychology, *Appl. Cognit. Psychol.* **27**: 360–372 (2013)
Published online 25 February 2013 in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/acp.2914

Quizzing in Middle-School Science: Successful Transfer Performance on Classroom Exams

MARK A. MCDANIEL*, RUTHANN C. THOMAS, POOJA K. AGARWAL, KATHLEEN B. MCDERMOTT and HENRY L. ROEDIGER

Washington University in St. Louis, St. Louis, MO, USA

Summary: We examined whether learning from quizzing arises from memorization of answers or fosters more complete understanding of the quizzed content. In middle-school science classes, we spaced three multiple-choice quizzes on content in a unit. In Experiment 1, the class exams included questions given on quizzes, transfer questions targeting the same content, and content that had not been quizzed (control content). The quizzing procedure was associated with significant learning benefits with large effect sizes and similar effect sizes for both transfer items and identical items. In Experiment 2, quiz questions focused on definitional information or application of the principle. Application questions increased exam performance for definitional-type questions and for different application questions. Definition questions did not confer benefits for application questions. Test-enhanced learning, in addition to other factors in the present quizzing protocol (repeated, spaced presentation of the content), may create deeper understanding that leads to certain types of transfer. Copyright © 2013 John Wiley & Sons, Ltd.

The use of summative testing to evaluate students' acquisition, retention, and transfer of instructed material is a fundamental aspect of educational practice and theory. However, a substantial literature has established that testing is not a neutral event—the act of retrieving answers to questions during testing can also enhance and modify memory for the tested information (Carpenter & DeLosh, 2006; Hogan & Kintsch, 1971; McDaniel & Masson, 1985; see Roediger & Karpicke, 2006a, for a review). Such findings suggest that educators might exploit testing as a technique to promote learning, not just as a way to assess learning. Converging on this suggestion, a number of quasi-experimental and correlational studies have demonstrated that quizzing can enhance performance on course assessments relative to no quizzing, for both online quizzing (Angus & Watson, 2009; Daniel & Broida, 2004; Kibble, 2007) and in-class quizzing (e.g., Bangert-Drowns, Kulik, & Kulik, 1991; Leeming, 2002; Lyle & Crawford, 2011). These patterns have been reinforced by recent experimental studies in college courses (McDaniel, Wildman, & Anderson, 2012) and middle-school courses (McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011; Roediger, Agarwal, McDaniel, & McDermott, 2011), showing significant improvement on course exams for material that has previously appeared on quizzes relative to material that has not been quizzed (for ease of exposition and in line with the literature, we will label this finding the *testing effect*).

One noteworthy limitation of nearly all laboratory and classroom experimental demonstrations of the testing effect is that the final exam questions have been the same as those used for the quizzes (e.g., Carpenter, Pashler, & Cepeda, 2009; McDaniel et al., 2011; Roediger, Agarwal, et al., 2011). In some educational contexts, providing identical questions on the quiz and the exam might be advocated when a large corpus of basic information and terms must be mastered, as in medical school (Larsen, Butler, & Roediger, 2008, 2009) or science courses (McDaniel et al., 2011). However, many educators and educational theorists would strongly object to including exam questions on initial quizzes (Popham, 2011). Accordingly, most extant experimental studies of the testing effect do not necessarily compel its broad utility in educational contexts.

Yet some recent laboratory work suggests that testing would benefit performance on exam items that are related but not identical to the items presented on the initial test (quiz). For ease of exposition and consistent with laboratory research (Rohrer, Taylor, & Sholar, 2010), we will label the benefits of initial testing on related but novel exam items as 'transfer'. According to this definition, testing would reinforce a representation of the tested concept that is flexible enough to be applied to different questions, which is typical of summative assessments used in educational contexts. We acknowledge that such transfer, if found, would be considered near transfer, as opposed to more extensive transfer, such as from the classroom to real-world settings, which might be considered an ideal consequence of educational instruction (cf. Barnett & Ceci, 2002).

Testing might be expected to produce very near transfer because testing improves learning and retention of associations relative to additional study of material (e.g., for learning the meaning associated with a new vocabulary item, see Karpicke & Roediger, 2008; for learning to associate pairs of words, see Carpenter, Pashler, & Vul, 2006). If acquired associations are bidirectional (A ↔ B), then initial testing in one direction (A—?) should improve performance on a novel final test for the reverse direction (B—?) relative to a study-only condition. Indeed, Rohrer et al. (2010) found that quizzing fourth and fifth graders to locate a particular county or city on a fictional map improved test performance on questions that required naming the county or city when given the location (and vice versa). Further, an experiment in a college course using online quizzing found similar benefits in associative transfer from quiz questions requiring generation of one element of a fact (e.g., for the quiz item 'All preganglionic axons, whether sympathetic or parasympathetic, release _____ as a neurotransmitter', in which 'acetylcholine' is the answer) to final test items requiring a

*Correspondence to: Mark A. McDaniel, Department of Psychology, Campus Box 1125, Washington University in St. Louis, St. Louis, MO 63130, USA.
E-mail: mmcdanie@artsci.wustl.edu

different element as the answer ('All _____ axons, whether sympathetic or parasympathetic, release acetylcholine as a neurotransmitter'; McDaniel, Anderson, Derbish, & Morrisette, 2007). In Experiment 1 of the present study, we examined whether quizzing in seventh-grade science classes would foster similar types of very near transfer in learning basic conceptual terms and their meanings, as indexed by the examinations used to evaluate the students (and assign grades).

Recent laboratory experiments also hint that testing might foster transfer in terms of application of target concepts beyond the context in which the concept was tested. For instance, Butler (2010) found that subjects given a cued recall test (with feedback) on concepts (e.g., wing structure for bats and birds) performed better on questions requiring transfer of those concepts to new contexts (e.g., wing structure for military aircraft) than did subjects who restudied the target concepts. Similarly, McDaniel, Howard, and Einstein (2009) reported that subjects required to recall technical passages (e.g., how brakes work) prior to rereading them received higher scores on later measures of inference and application questions compared with subjects who reread the passage without intervening recall. Likewise, Johnson and Mayer (2009) found that subjects given a practice test about lightning formation after a multimedia presentation performed better on novel questions requiring transfer of the tested concepts than did subjects who rewatched the multimedia presentation without a practice test. These findings imply that testing can also promote better understanding of constructs, perhaps including a more organized (Zaromb & Roediger, 2010) or detailed mental model of the target information.

Accordingly, quizzing in the classroom might promote deeper or more complete learning of the target concepts, such that performance on exam questions that required application of these concepts would be enhanced relative to no quizzing. The idea here is that initial quizzing on target concepts (e.g., competition for resources) might promote performance (relative to no quizzing) on novel exam items requiring application (e.g., 'Both foxes and raccoons on Long Island eat pheasant, which in recent years, has been in decline. The foxes and raccoons' situation is an example of what ecological process?'). To examine this possibility, in Experiment 2, we investigated the extent to which several types of quiz items (definitional or application) would enhance performance on exam items that required application of target concepts (as illustrated earlier) in a context that was different from that seen in the quizzes. We emphasize that the purpose of this study was to investigate the degree to which quizzing (testing) would improve classroom examination performance on several types of exam questions (definitional and application) commonly used in middle-school classrooms, when the quiz items were not identical to exam items.

EXPERIMENT 1

The learning of basic conceptual terms and their meanings is one key educational objective in the middle-school science classes that were participating in our studies. That is, exams assess students' knowledge of definitions of core concepts, with multiple-choice questions that prompt students to either provide a definition of a concept (e.g., What are amino acids?) or provide the term for the concept given a definition (e.g., What are the smaller units that form proteins?). In our experiments, we used the actual content and exams used by the teacher in the course, not artificial materials. The present experiment extended the McDaniel et al. (2011) study by including conditions in which the type of question changed from quiz to exam. Specifically, for quiz items that provided the concept term in the stem and required a definition for the response (for ease of exposition, we term these *definition-response* questions), the exam items provided the definition in the stem and required the concept term for the response (we term these *term-response* questions). As an example, a definition-response quiz question would be as follows:

What of the following correctly describes active transport?

A. When a cell moves water without the use of energy.
B. The movement of RNA from the Golgi body to the nucleus.
C. The transportation of DNA from the endoplasmic reticulum to the nucleus.
D. The movement of material through the cell membrane using energy.

The subsequent term-response exam question would then be as follows:

What process is used when a cell needs to take in a substance that is higher in concentration inside the cell then outside and requires the cell to use energy to complete this process?

A. Passive transport
B. Active transport
C. Osmosis
D. Diffusion

In a parallel vein, term-response quiz items preceded definition-response exam items. In contrast to previous studies in which the final test questions were identical to the quiz questions (e.g., Carpenter et al., 2009; McDaniel et al., 2011; Roediger, Agarwal, et al., 2011), superficial learning of a particular response *per se* from practice on quizzes would not be sufficient to support performance on the exam questions. Thus, in the present paradigm, unlike in the work of McDaniel et al. (2011) and Roediger, Agarwal, et al. (2011), students must acquire an integrated representation of the definition with the concept term, a representation that is perhaps somewhat gist based (in comprehension theories, a more propositional than verbatim-level representation; Kintsch, 1988), for the quiz item to enhance performance on the different-exam item.

If learning via testing and feedback (i.e., via answering 'closed-book' questions on quizzes) can be used flexibly as some earlier research suggests, one straightforward prediction in the current experiment is that quizzing with definition-response questions will enhance performance on term-response exam items. Similarly, quizzing with term-response questions will enhance performance on the definition-response exam items, relative to no quizzing. A related issue concerns whether these transfer effects of testing, if obtained, would be as robust as when the exam items were identical to the quiz items.

Rohrer et al. (2010) found that testing produced associative transfer effects that were at least as sizable as effects for identical items. To gauge the magnitude of any near-transfer effects due to quizzing, we also included identical items on the quiz and exam assessments.

Alternatively, it is possible that in a classroom setting, quizzing will not benefit performance on different test items, especially where multiple-choice testing is relied upon. In this regard, it is important to note that the robust associative transfer effects in the works of both McDaniel et al. (2007) and Rohrer et al. (2010) were produced by cued recall (short-answer) quizzes. In the present classroom context, the quizzes used multiple-choice questions because they were administered in class through interactive response systems (clickers). Multiple-choice quizzing effects are generally less robust than short-answer quizzing effects (Kang, McDermott, & Roediger, 2007; McDaniel et al., 2007; but see McDaniel et al., 2012, for an exception). So that a strong multiple-choice quizzing procedure to test these possible outcomes could be provided, quizzes were repeated three times over the course of the target science units.

Method

Participants

One hundred forty-two seventh-grade science students from a public middle school located in a suburban middle-class community in the Midwest participated in this study. Parents were informed of the study, and written assent from each participant was obtained in accordance with the Human Research Protection Office. The school board, principal, and teachers agreed to participate in the study. Three students declined to have their data included.

Materials and design

A 3 (initial quiz question type: definition response, term response, and nonquizzed) $\times$ 2 (exam question type: definition response and term response) within-subjects design was used. There were three initial quiz phases: pre-lesson (before the teacher's lesson, but after reading an assigned chapter from the textbook, assuming students followed the teacher's instructions), post-lesson (after the teacher's lesson), and review (24 hours before the exam). The exam was administered at the end of the unit, 11 days (on average) after the material was first introduced (Figure 1).

Before continuing, it is worthwhile to comment on the limitations of the design. As guests in the classrooms, we were unable to substantially alter the normal classroom practices. Accordingly, we could not include nonquizzed conditions that would equate total exposure and spacing of the material to that provided by the quizzing conditions. Doing so would have required the teacher to significantly change her lesson plans and daily lecture content. Further, including nonquizzed exposure conditions that paralleled the quiz conditions would have significantly reduced the number of exam items (observations) in each experimental condition (refer to the following discussions for these details). Unlike in the laboratory, the amount of target material could not be increased to accommodate additional within-subjects conditions. Thus, we emphasize that the interpretation of any benefits associated with quizzing (if found) in the present study is limited by these constraints imposed from integrating the quizzing into ongoing instruction delivered in an authentic classroom. We return to these interpretational issues in the discussion of the results.

Material from three science units (cells, machines/energy, and animals) was used in this study. Across the three units, 120 multiple-choice questions with four alternatives (one correct answer and three lures) were developed by the experimenters and the teacher. Forty items were initially quizzed (on pre-lesson, post-lesson, and review quizzes) in a definition-response format, 40 items were initially quizzed in a term-response format, and 40 items were not quizzed. At the end of each unit, students received definition-response

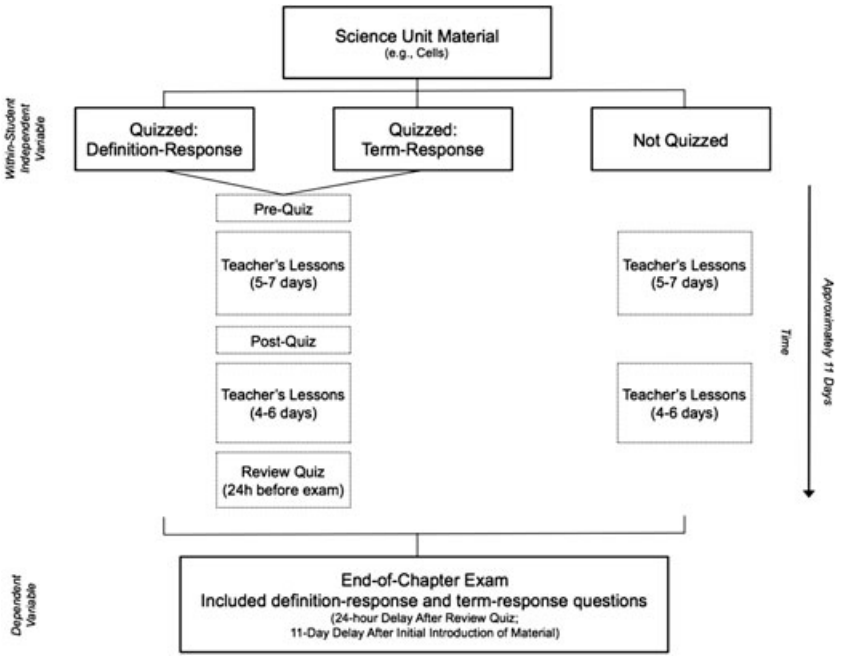

Figure 1. Experimental procedure for Experiment 1

questions on half of the items and term-response questions on half of the items, such that across units, 20 items were in each of the six conditions formed by the 3×2 factorial of quiz status (definition-response quiz item, term-response quiz item, and nonquizzed item) and type of exam question (definition response and concept term). No other questions appeared on the exams; that is, the exams consisted entirely of the multiple-choice questions used in the experiment. All target facts were covered in assigned readings and the teacher's lessons. Each of the six classroom sections ($M = 24$ students) had a different random assignment of items to the six conditions.

Items were quizzed in the same format for the pre-lesson, post-lesson, and review quizzes. Across these three initial quizzes, the wording of the question stems remained the same, but the order of the four alternative answers was randomly ordered for each quiz. Questions were based on the definition provided in the required reading and reviewed during classroom lectures. Definition-response questions included the concept term in the stem, and students had to select the correct definition from the alternatives. For term-response questions, the definition was given in the question stem, and students had to select the correct term response from the alternatives. Question stems on the exam were slightly reworded so that none of the questions from the initial quizzes was identical to the exam.

For term-response questions, the incorrect multiple-choice alternatives (i.e., lures) were plausible concept terms, including other concepts from the unit as well as plausible concepts not covered in the class. For example, the alternatives for the term-response quiz question about active transport (Appendix) included the lures *passive transport*, *osmosis*, and *diffusion*, all of which were concepts covered in the unit that served as correct answers to other term-response questions. The alternatives for the term-response quiz question about passive transport included the lures *active transport*, *energy conservation transport*, and *cell transport*, with only *active transport* being a concept covered in the class. For definition-response questions, the lures were plausible definitions, which included correct definitions for other concepts covered in the unit as well as other plausible definitions not covered in the class (e.g., the correct definition with several features changed to make it incorrect; e.g., see lure 'C' for the question 'What is the definition of temperature?' in the Appendix). Note that for quiz and exam questions of the same type, the four alternative answers were held constant across quiz and exam questions. The Appendix displays examples of definition-response and term-response questions for quizzes and exams.

Procedure

An experimenter administered all initial quizzes (pre-lesson, post-lesson, and review) via a clicker response system (Ward, 2007), and items on initial quizzes were presented in the same order as presented during the lessons. Pre-lesson quizzes were administered after students read an assigned chapter from the textbook, but before the teacher discussed the information (Figure 1 provides a timeline of the activities in the procedure). Thus, pre-lesson quiz performances reflected students' learning from the assigned reading, their preexisting knowledge about the topic, or both. Scores on the pre-lesson quizzes did not count toward students' final grades. The teacher was not present for these quizzes to avoid potential bias toward particular items during her lesson, which immediately followed the pre-lesson quiz. That is, the teacher was unaware which content was assigned to which condition for any individual class.

For each item, the question and four multiple-choice alternatives were displayed on a large projection screen at the front of the classroom and were concurrently read aloud by the experimenter. Students were required to respond to each question by pressing the A, B, C, or D button on their individual clicker remotes. After all students responded, a green checkmark appeared next to the correct response, and 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 pre-lesson 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 individual score by their assigned clicker number but were not aware of other students' clicker numbers. The teacher then proceeded with the lesson.

Post-lesson quizzes were administered after the teacher had covered all material for a particular chapter. Review quizzes were administered 24 hours before exams. Overall, the procedure for post-lesson and review quizzes was identical to pre-lesson quizzes with two exceptions. The teacher was present during these quizzes, and in compliance with the teacher's course format, scores from all of the post-lesson and review quizzes in the course counted in total for a small portion (approximately 10%) of each student's cumulative grade. Additionally, students were not explicitly told when post-lesson quizzes would be administered, but they were aware that the review quiz was the day before the exam. It should be noted that depending on the time available, during a class prior to the exam, the teacher usually would review all materials (quizzed and nonquizzed). Accordingly, the recency of exposure (relative to the exam) for quizzed and nonquizzed material was equated for some (although not all) of the units.

The teacher administered exams using paper and pencil. The scores on these exams accounted for 50% of each student's overall grade. The students were informed of their overall score 1–2 days after the exam, but they did not receive corrective feedback on an item-by-item basis. The teacher's typical lesson plans remained unchanged throughout the procedure. Students were exposed to all of the information contained on the exam via the teacher's lessons, homework, and worksheets; therefore, students were exposed at least once to nonquizzed items during typical classroom activities.

Results

The data from 40 students who qualified for special education ($N = 8$), directed studies ($N = 14$; students who do not qualify for special education services but who receive additional classroom support), or gifted programs ($N = 18$) were excluded from the analyses. Further, 38 students who did not complete all of the required quizzes or exams were also excluded. The final sample consisted of 61 seventh-grade students. However, the pattern of results and statistical

outcomes remained the same when data from the 38 absent students and the 40 other students in special programs were included in the analyses. On average, for the entire sample, quizzes improved term-response exam performance by 12–15%, which was a bit larger effect than that obtained for the final sample of 61 students; and quizzes improved definition-response exam performance by 9–10%, which was a bit smaller effect than that obtained for the final sample (see Figure 2 for the means from the final sample). Given that our primary interest was in repeated testing and transfer effects, the analyses reported in the following collapsed across the specific units. For all analyses, the significance level was set at an alpha level of .05.

Initial quiz performance

Performance on the initial quizzes was analyzed using a 3×2 repeated-measures analysis of variance with initial quiz placement (pre-lesson, post-lesson, and review) and quiz question type (definition response and term response) as within-subject variables, collapsed across the three science units. As shown in Table 1, student accuracy increased from the pre-lesson quizzes ($M = 0.48$) to the post-lesson ($M = 0.71$) and review ($M = 0.80$) quizzes for both definition-response and term-response questions, as revealed by a main effect of quiz placement, $F(2, 120) = 505.55$, $MS_e = 0.006$, $\eta_p^2 = .89$. Also, in general, students performed better on term-response questions ($M = 0.68$) than on definition-response questions ($M = 0.65$), $F(1, 60) = 5.53$, $MS_e = 0.011$, $\eta_p^2 = .08$. This effect was qualified by a significant interaction with quiz placement, $F(2, 120) = 5.79$, $MS_e = 0.005$, $\eta_p^2 = .09$. Performance on pre-lesson quizzes did not differ for definition-response and term-response questions, $F < 1$. For the post-lesson and review quizzes, however, performance was better for term-response questions than for definition-response questions, $F(1, 60) = 7.91$, $d = 0.32$, and $F(1, 60) = 13.48$, $d = 0.40$, respectively.

Exam performance

Performance on the exam was analyzed using a 3×2 within-subjects analysis of variance with initial quiz question type (nonquizzed, definition response, and term response) and exam question type (definition response and term response) as variables (see Figure 2 for means). Exam performance on term-response questions ($M = 0.78$) was significantly higher than for definition-response questions ($M = 0.75$), $F(1, 60) = 17.65$, $MS_e = 0.008$, $\eta_p^2 = .23$. More importantly, there was a significant main effect of quiz question type, $F(2, 120) = 82.97$, $MS_e = 0.010$, $\eta_p^2 = .58$, such that exam performance was enhanced when the target content had been previously quizzed ($M = 0.82$ for term-response questions and $M = 0.80$ for definition-response questions) than when it had not been quizzed ($M = 0.67$). Further, a significant interaction between quiz question type and exam question type, $F(2, 120) = 5.93$, $MS_e = 0.009$, $\eta_p^2 = .09$, indicated that the magnitude of performance benefits from prior quizzing (relative to no quizzing) differed between definition-response and term-response exam questions. Planned comparisons confirmed this interpretation. In particular, the difference in exam scores for quizzed compared with nonquizzed questions was greater for definition-response questions ($M = 0.17$) than for term-response questions ($M = 0.10$), $F(1, 120) = 16.51$, $MS_e = 0.009$, $d = 0.56$.

Table 1. Initial quiz performance (proportion correct) as a function of quiz placement and question type for Experiments 1 and 2

	Initial quiz placement		
	Pre-lesson	Post-lesson	Review
Experiment 1 ($n = 61$)			
Term response	0.48 (0.01)	0.73 (0.01)	0.82 (0.01)
Definition response	0.49 (0.01)	0.69 (0.02)	0.78 (0.02)
Experiment 2 ($n = 90$)			
Term response	0.56 (0.02)	0.82 (0.02)	0.85 (0.02)
Application	0.51 (0.02)	0.79 (0.02)	0.86 (0.02)

Note. Standard error is noted in parentheses.

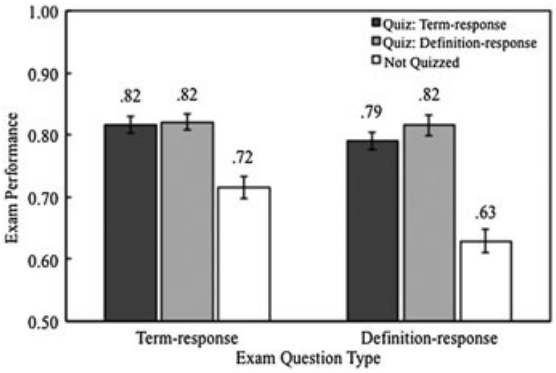

Figure 2. Exam performance (proportion correct) on term-response and definition-response questions as a function of initial quiz question type in Experiment 1. Error bars represent standard error of the mean

One critical aim of this study was to explore whether quizzing would promote near transfer of target content, that is, benefits even when the quiz and exam questions were different types. Planned comparisons revealed that quizzing improved exam performance for both same-type and different-type questions (relative to no quizzing). Students scored higher on term-response exam questions after being quizzed with either corresponding term-response questions or definition-response questions compared with nonquizzed term-response questions, $F(1, 120) = 35.26$, $MS_e = .009$, $d = 0.81$, and $F(1, 120) = 38.08$, $MS_e = 0.009$, $d = 0.86$, respectively. Similarly, students scored higher on definition-response exam questions after being quizzed with either corresponding definition-response questions or term-response questions compared with nonquizzed questions, $F(1, 120) = 118.51$, $MS_e = 0.009$, $d = 1.33$, and $F(1, 120) = 87.84$, $MS_e = 0.009$, $d = 1.23$, respectively. Importantly, the levels of exam performance following quizzing were virtually identical when the quiz and exam question types were the same (term-response to term-response items and definition-response to definition-response items; $M = 0.82$) and when they were different (term-response to definition-response items and definition-response to term-response items; $M = 0.81$; $F < 1$).

Discussion

The most important finding of this experiment was that spaced quizzing with feedback promoted flexible use of

target content on later exam performance. Especially telling was that the effect sizes for the different-item and same-item conditions were comparable (for exam items for which the definition had to be provided, $d = 1.33$ and 1.23 for identical and different-item performances, respectively; for exam items for which the concept term had to be provided, $d = 0.81$ and 0.86, respectively) even though the same-item conditions had the advantage that the response alternatives for each question were repeated across the quiz and exam items (thus allowing correct responding to be potentially mediated by memory of the answer *per se*). This pattern implies that spaced quizzing reinforces learning of the association between the term response and its definition in a way that enables near transfer to new questions about the meaning of the concept (from the term response to the definition and the definition to the term response). The present findings extend several existing laboratory studies that reported that quizzing (testing) promoted associative transfer on final tests. Carpenter et al. (2006) quizzed associations between pairs of words, and Rohrer et al. (2010) tested associations between locations on fictional maps and city (or region) names; both reflect basic paired-associate material, with little inherent meaning. The current materials were more complex, targeting the meaning of basic science concepts. Accordingly, the present findings show that spaced quizzing (with feedback) enhances learning of a unified concept that is imbued with meaning (rather than arbitrary pairings such as words or locations and names).

As important, the present results are the first to demonstrate that a particular quizzing procedure, which includes other potentially effective components such as feedback and spacing, enhances performance on near-transfer items relative to no quizzing in an authentic middle-school classroom. Our dependent measures were the exams that were used to assign grades to the students. As such, students were presumably motivated to learn the target material, and this material (quizzed and nonquizzed) was focused on in class lectures, in the textbook, in homework assignments, and in pre-exam classroom reviews. Nevertheless, material that was quizzed was better learned than nonquizzed material. It is also noteworthy that these exams were representative of the exams used in previous years in the science class, and they reflected the learning objectives, at least in part, of the middle-school science course. That is, acquisition of the content favored by the present quizzing procedure was central to the course objectives and is content that is likely incorporated into course objectives in many middle-school science classes. From this perspective, the benefits of the quizzing procedure were impressive. Performance levels on material that was not quizzed were at the D level for the school's grading scale for term-response questions and the F level for the definition-response questions; quizzing generally increased performance levels on the material to a C/C+ level (students' semester grades were also determined by classroom projects, homework, and other assignments, so that overall semester grades would be higher).

Clearly, the nature of the learning stimulated by the quizzing procedure cannot be unequivocally determined. For these types of authentic exam items, performance could be supported by committing the key concept terms and their definitions to memory, as well as by a richer understanding of the concepts. Quizzing with feedback might have promoted either or both. Further, as noted earlier, the benefits associated with quizzing and feedback could have been produced by more exposure to the quizzed than the nonquizzed material, to more spacing of the quizzed material, or to more recent exposure (relative to the exam) of the quizzed material (although as noted in the procedure, for some of the units, nonquizzed material was also reviewed prior to the exam). However, we cautiously suggest that trying to improve memory of the concepts through re-presentation *per se* (i.e., restudy instead of quizzing) would not have conferred the same benefit as quizzing. A parallel experiment in the same middle school found that quizzing (with feedback) but not re-presentation of the material benefited exam performance when the exam items were identical to the quiz items and the re-presented material (Roediger, Agarwal, et al., 2011, Experiment 2). (Also, considerable laboratory evidence has shown that testing with feedback boosts later retention more so than repeated study, especially when the final test is delayed rather than immediate; see Roediger & Karpicke, 2006a, for a review.) In the present experiment, the quizzing (with feedback) effects were as large when the exam items were different from the quiz items as when the exam items were identical to the quiz items. To the extent that the Roediger, Agarwal, et al. (2011) results generalize to the materials used here, it thus seems likely that re-presentation *per se* of the target concepts would have failed to increase exam performance to the level produced by quizzing, even in the different-exam-item conditions. Of course, decisive evidence for the benefits of quizzing (with feedback) *per se* would require directly pitting a re-presentation condition against quizzing with the current target material and learning environment.

EXPERIMENT 2

The exams in Experiment 1 reflected a straightforward reproduction of the scientific concept terms and their definitions targeted in the lessons. Another potential learning objective, however, would be for students to understand the instantiation of these concepts in a variety of contexts. One initial laboratory study with education-like materials has demonstrated that testing (quizzing) of concepts illustrated in one context can enhance transfer to final assessments that test the understanding or application of that concept in another context (Butler, 2010). For example, in Butler (2010), an initial test question on a passage about bats was 'A bat has a very different wing structure from a bird. What is the wing structure of a bat like relative to a bird?' These questions (with feedback) produced better performance on a final question (than did restudy) that required application of the concept in a military context ('The U.S. Military is looking at bat wings for inspiration in developing a new type of aircraft. How would this new type of aircraft differ from traditional aircrafts like fighter jets?').

Some educators might consider the preceding laboratory instantiation of transfer as somewhat contrived, as answering could simply involve recalling the structure of a bat wing and describing how that would look on an aircraft.

Accordingly, a major purpose of Experiment 2 was to examine how quizzing might impact performance on exam questions that required students to figure out what principle or construct was being illustrated in a particular scenario or situation (we label these *application* questions; illustrated in the Appendix)—rather than relying on a concept explicitly mentioned in the question (e.g., bat-wing structure as in Butler, 2010) to describe the target situation.

We reasoned that two kinds of quiz items could potentially enhance performance on application exam questions (relative to no quizzing). One hypothesis is that using quiz application questions, along with feedback, may stimulate students to consider the target principle and to appreciate how it operates. Forcing this kind of deep processing during quizzing plausibly would facilitate performance on exam application questions (ones that instantiate the principle in a different context from that seen at the quiz). (See the Appendix for examples of the quiz and exam application questions.) The theoretical idea is that the application quiz question forces students to consider the implications of the principle and how it is instantiated, thereby supporting the generation of inferences (or abstraction) required by the exam question (in line with the transfer-appropriate processing principle from the memory literature; McDaniel, Friedman, & Bourne, 1978; Morris, Bransford, & Franks, 1977; Roediger, Weldon, & Challis, 1989). Preliminary support for this idea is a recent finding showing quiz-related gains on inference exam questions in a college class (Glass, 2009).

A second hypothesis is that quiz items focusing on definitions stimulate better learning of the concept and thereby facilitate its application. Experiment 1 showed that spaced quiz questions (with feedback) that required students to match a definition with the correct concept term promoted more durable learning of the concept term and its definition (see also McDaniel et al., 2011, for quiz-enhanced retention in middle-school classes of up to 9 months). Better-learned definitions of target concepts may provide students a more accurate foundation for discerning the particular concept illustrated in application questions on a later test. If so, then term-response quiz questions (the definition was provided in the question stem and students chose the correct concept term from the answer options) might increase performance on the application exam items relative to the no-quiz control.

Another interesting possible sequence for quizzing is one in which application quiz items are followed by exam items that focus on definitional information (like the questions in Experiment 1). We were able to examine this particular transfer sequence in Experiment 2 because exams in the science classes included such term-response questions. We reasoned that students attempting to answer application quiz items might also retrieve or activate definitional information for the target principles, thereby improving learning of that information. For instance, the idea is that the application quiz item on 'competition' (the 'foxes and raccoons compete for pheasant' question; Appendix) would stimulate consideration of the definition of 'competition', which in turn could enhance retention of the definition, a better understanding of the definition, or both. Thus, we thought it possible that application quiz items (with feedback) might promote performance on the term-response exam items relative to a no-quiz control.

Method

Participants

One hundred forty-two eighth-grade science students from the same public middle school located in a suburban middle-class community in the Midwest participated in this experiment. Parents were informed of the study, and written assent from each participant was obtained in accordance with the Human Research Protection Office. The school board, the principal, and the teachers agreed to participate in the study. Three students declined to have their data included. Students in Experiment 2 did not participate in Experiment 1.

Materials and design

A 3 (initial quiz question type: term response, application, and nonquizzed) $\times$ 2 (exam question type: term response and application) within-subjects design was used. Similar to Experiment 1, there were three initial quiz phases: pre-lesson, post-lesson, and review. Retention was measured at the end of the unit ($M = 16$ days after material was first introduced).

Materials from two science units (ecology and environment) were used in this study. Terms were chosen for inclusion in this experiment only if the term could be applied in different contexts. For example, erosion has many different contextual applications (streams, oceans, wind, etc.). Across the two units, the experimenters and the teacher developed 30 multiple-choice questions with four alternatives. Ten questions were initially quizzed (on pre-lesson, post-lesson, and review quizzes) in a term-response format, 10 questions were initially quizzed in an application format, and 10 questions were not quizzed. At the end of each unit, students received term-response questions on half of the items and application questions on half of the items, such that across units, five items were in each of the six conditions. The exams for ecology and environment included an additional 29 and 27 questions (mix of multiple-choice, short-answer, fill-in-the-blank, and graph/diagram questions), respectively, which were related to unit material, but not quizzed or assessed for experimental purposes. All target facts were covered in assigned readings and the teacher's lessons. Each of the six classroom sections ($M = 24$ students) had a different random assignment of items to the six conditions.

Items were quizzed in the same format for the pre-lesson, post-lesson, and review quizzes. Across these three initial quizzes, the wording of the question stems remained the same, but the order of the four alternative answers was randomly ordered for each quiz. The same four multiple-choice alternatives were then used on exam items, but the presentation of the alternatives was randomly ordered for each quiz and exam. This design feature was implemented so that familiarity (familiarity produced by repetition of items across quizzes) could not be used to guide selection of the correct response on the exam. Our reasoning was that by repeating both lures and the correct response for all quizzes and the exam, familiarity would accrue equally to all possible response alternatives.

Another important design feature was that 35% of the lures were plausible concept terms, and 65% of the lures were concepts tested in the same unit (and, thus, also correct answers on other questions). Accordingly, simply

remembering what concept terms were answers to the quiz questions could not ensure correct response on exam items (except for the term-response quiz–term-response exam items). The Appendix displays examples of term-response and application questions with the multiple-choice alternatives used for quizzes and the exam in this experiment. Question stems on the exam were reworded (at least slightly for term-response quiz–term-response exam pairings) so that none of the questions from the initial quizzes was identical to the exam questions.

Procedure

Procedures in Experiment 2 were identical to those reported in Experiment 1. Briefly, an experimenter administered all initial quizzes (pre-lesson, post-lesson, and review) in the classroom using a clicker response system. Quiz questions were presented in multiple-choice format with corrective feedback provided after all students responded. Scores from pre-lesson quizzes did not count toward each student's overall grade. Post-lesson and review quizzes counted for a small portion (10%) of each student's cumulative grade. Exams were administered using paper and pencil. Scores on these exams accounted for 50% of each student's overall grades.

Results

The data from 18 students who qualified for special education ($N = 9$) or gifted programs ($N = 9$) and 31 students who did not complete all of the required quizzes or exams were excluded from the analyses. The final sample for analyses consisted of 90 eighth-grade students. As in Experiment 1, the pattern of results and statistical outcomes remained the same when data from the 18 students in special programs and 31 absent students were included in analyses. On average, for the entire sample, quizzes improved term-response exam performance by 5–8%, which was a bit smaller effect than obtained for the final sample of 90 students; and quizzes improved appllication-question exam performance by 2–5%, which was nearly identical to that obtained for the final sample (see Figure 3 for the means from the final sample). The analyses collapsed across the unit or course topic. All results were significant at an alpha level of .05 unless otherwise noted. The plan for analyses was identical to Experiment 1.

Initial quiz performance

Table 1 displays students' mean performance on initial quizzes as a function of initial quiz question type and initial quiz placement. Overall, students' performance increased from the pre-lesson (53%) to the post-lesson (81%) and review (86%) quizzes for both term-response and application questions, as revealed by a main effect of quiz placement, $F(2, 178) = 305.78$, $MS_e = 0.018$, $\eta_p^2 = .78$. Also, there was a trend for students' performance to be better on term-response questions (74%) than on application questions (72%), $F(1, 89) = 2.96$, $MS_e = 0.024$, $p = .09$. The interaction between quiz placement and initial question type was not significant, $F(2, 178) = 2.23$, $MS_e = 0.015$, $p = .11$.

Exam performance

Figure 3 displays students' mean performance on term-response and application exam questions as a function of initial quiz question type. Exam performance was significantly higher on term-response questions than on application exam questions, $F(1, 89) = 5.34$, $MS_e = 0.024$, $\eta_p^2 = .06$. More importantly, there was a significant main effect of quiz question type, $F(2, 178) = 9.95$, $MS_e = 0.028$, $\eta_p^2 = .10$, such that exam performance was enhanced when the target content had been previously quizzed ($M = 0.83$ for term-response questions and $M = 0.83$ for application questions) relative to nonquizzed content ($M = 0.76$). A significant interaction between quiz question type and exam question type, $F(2, 178) = 4.79$, $MS_e = 0.034$, $\eta_p^2 = .05$, indicated that the magnitude of performance benefits from prior quizzing (relative to no quizzing) was greater for term-response exam questions than for application exam questions. A direct comparison confirmed this interpretation: The quizzing effect was greater for term-response ($M = 0.10$) exam questions than for application exam questions ($M = 0.04$), $F(1, 178) = 5.303$, $MS_e = 0.034$, $d = 0.28$.

Planned comparisons were conducted to explore whether quizzing promoted benefits on application exam questions. Students scored higher on application exam questions after being quizzed with application questions compared with nonquizzed questions, $F(1, 178) = 4.77$, $MS_e = 0.034$, $d = 0.27$. However, students did not score any higher on application exam questions when they were quizzed with related term-response questions compared with the nonquizzed questions ($F < 1$). On term-response exam questions, as in Experiment 1, students scored higher after term-response quiz questions compared with no-quiz questions, $F(1, 178) = 21.35$, $MS_e = 0.034$, $d = 0.56$; importantly, for present purposes, students also scored higher when they had been quizzed with application questions compared with nonquizzed questions, $F(1, 178) = 7.65$, $MS_e = 0.034$, $d = 0.34$.

Discussion

Experiment 2 extended the results from Experiment 1 by demonstrating not only that a quizzing procedure on definitional content that incorporates repeated, spaced quizzes

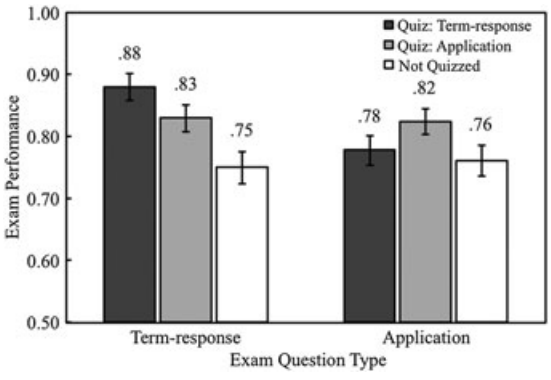

Figure 3. Exam performance (proportion correct) on term-response and application questions as a function of initial quiz question type in Experiment 2. Error bars represent standard error of the mean

(with feedback) can promote learning of that content but also that using the procedure for quizzing applications can promote learning for target scientific principles. One benefit of application quizzes was evidenced when the exam questions were definitional. Several interpretations of this pattern are possible. One general type of explanation already noted is that the application quiz item with feedback provided additional spaced exposure to the concept and further provided exposure that was possibly more recent to the exam than nonquizzed material (although not always, as specified in Experiment 1). This repeated exposure to application of the concept could have improved learning (or memory) of the concept. In the general discussion, we return to this explanation and also consider possible indirect effects of quizzing that may have contributed to its benefits.

Another possible interpretation for the significant benefits of application quizzes to term-response exam questions rests on the benefits produced by the active processing required by the quizzes (rather than mere exposure to content).[1] More specifically, when students were confronted with application quiz items, they may have retrieved definitional information to support application of the concept. Retrieval of this definitional information during quizzing would be expected to promote retention that supports performance on term-response exam questions. Or quizzing could have enhanced memory for the applications themselves; then when students were confronted with the term-response questions on the exam, they used the illustrative applications to infer definitions of the concepts.

Another benefit associated with the application quizzing procedure was that it enhanced performance on application exam items for the target principles relative to no quizzing. This pattern possibly represents transfer because the context of the application in the quiz items was different from that encountered on the exam questions. For example, the context for application questions on *competition* changed from foxes and raccoons competing for pheasant on the quiz question to groups of pandas competing for bamboo on the exam (Appendix). Answering application quiz questions may have stimulated students to deeply process the target principle (i.e., competition for limited resources) and actively consider its implications for the application context encountered on the exam. Such processing would presumably be beneficial for the application exam question.

[1] A prosaic interpretation of benefits of the quizzes is that the answers students learned for the quiz items could be selected as the answers to the corresponding exam items (the correct answers were identical for application quiz questions and the corresponding term-response questions focusing on definitional information, and the lures were also identical across corresponding quiz and exam questions, which might have served as additional cues for identifying the correct exam answer). Weighing against this possibility, the correct-answer response for a particular question could also be a lure for questions about a different concept (65% of the concept terms in Experiment 2 served both as incorrect lures and correct answers), thereby precluding extensive reliance on learning particular answers as a basis for responding on the exam. Further, and critically, if the benefit of application quizzing was merely a consequence of rote learning of particular answers, then one would expect that the term-response quiz items would consistently provide similar benefits to application exam items. However, term-response quiz items produced no significant benefit on the application exam items (relative to no quizzing).

GENERAL DISCUSSION

The current experiments suggest that a spaced testing (quizzing) procedure, along with feedback, enhances knowledge that can be flexibly used for different test items appearing on later exams. In previous experiments conducted in middle-school classes, the items on the exams were identical to those presented on the quizzes (McDaniel et al., 2011; Roediger, Agarwal, et al., 2011). In these cases, the benefits for exam performance could rest simply on retrieval practice of particular answers during quizzing. If the match between quiz and exam questions were a boundary condition for achieving robust benefits of quizzing (with feedback), then its applicability in the classroom would be limited for a number of reasons, including the observation that many instructors would not find it acceptable to use the identical question on the quizzes and the exam (e.g., Mayer et al., 2009). Further, quizzing would only be useful if the precise nature of the exam questions were known ahead of time (Rohrer et al., 2010).

The present results showed that a spaced quizzing procedure with feedback can promote learning that is deeper than just retaining a particular answer (i.e., providing an identical answer on a summative exam); rather, learning can be transferred to new forms of assessment (i.e., to new exam items). Experiment 1 showed that the quizzing procedure improved performance on different exam items requiring a reverse association between a term response and its definition from the form of the item that was quizzed. Especially, large effect sizes were observed for the quizzing effects when the exam items required a definition as a response, and this effect was produced even after the quizzes had contained different items (ones requiring the term response as a response).

Experiment 2 further showed that the quizzing procedure promoted transfer from applying a concept in a concrete context on the quiz to better retention of definitional information on a later exam. Importantly, answering the application questions generally also permitted transfer to applying the principle in a new context. Thus, the present spaced quizzing procedure implemented in the classroom, particularly with application questions, encourages learning that is richer than might be expected from simple memorization; application questions produced significant positive effects on both term-response exam items and different application exam items. We emphasize that if the current effects were based on simple memorization of a repeated correct answer from among repeated lures, then the effects of term-response (definitional) quiz items on application exam items should have been as robust as the effects of application quiz items on term-response exam items. Experiment 2 convincingly counters this possibility, in part because the baseline levels (no quiz) for the application and definition-response exam items are comparable (Figure 3). We suggest that quizzing might benefit transfer even more robustly by varying the contexts used across the three application quizzes so that the target concept/principle is more broadly and completely illustrated (see Glass, 2009, for a related finding). We note that it is possible that requiring middle-school students to repeatedly study applications might produce transfer effects similar to those found here for the application quiz items. It is worth

mentioning, however, that in laboratory experiments, repeated study has not produced transfer to new applications, whereas quizzing (with feedback) has (Butler, 2010).

The present findings appear to extend the limited laboratory work indicating that initial tests can produce transfer on novel questions (Butler, 2010; Carpenter et al., 2006; Chan, McDermott, & Roediger, 2006; Johnson & Mayer, 2009; Rohrer et al., 2010). One laboratory study showed that initial testing promoted transfer of conceptual information to new contexts and applications (Butler, 2010) but left uncertain whether the finding would generalize to content from classrooms, to exams that determine students' grades, and to long-term retention that is required in classroom settings. The present experiments suggest that quizzes (in combination with other components such as spacing and feedback), especially application quizzes (Experiment 2), can promote flexible application of target content to new contexts (i.e., not previously instructed) in authentic classroom settings. In these classrooms, unlike the laboratory, the target material was emphasized in class lectures and demonstrations and was reinforced with assigned reading, homework, and in-class reviews. Yet spaced quizzing (with feedback) still was associated with performance gains on exams with novel questions.

The current study also extends the laboratory testing experiments by showing benefits on novel exam items after multiple-choice quizzing. Prior laboratory experiments used recall tests (for both initial and final testing), which are considered to be more demanding in terms of retrieval than multiple-choice tests. On some accounts, the retrieval difficulty of the initial test items determines the potency of the testing effect (Gardiner, Craik, & Bleasdale, 1973; McDaniel & Masson, 1985; Pyc & Rawson, 2009; Rohrer et al., 2010). Accordingly, some laboratory experiments left open the possibility that transfer would be limited to instances in which initial testing required retrieval challenges (cf. Rohrer et al., 2010). The present experiments establish that spaced multiple-choice quizzes with feedback may be sufficient to enhance performance on novel test items. One factor that might be important in this regard is that the format on the initial quizzes matched that on the exams in both the current experiments (multiple-choice items on both quizzes and exams) and the laboratory studies (cued recall on initial and final tests; Butler, 2010; Carpenter et al., 2006; Chan et al., 2006; Rohrer et al., 2010; see McDaniel et al., 2007, for an exception).

As mentioned throughout, other factors included in the present protocol (in addition to, or instead of, retrieval required by the quiz) could play a role in the effects, such as repeating quizzes (cf. McDaniel et al., 2012) or the presence of correct answer feedback. Learning from feedback can itself be beneficial (e.g., Butler & Roediger, 2008; McDaniel & Fisher, 1991), especially when a correct answer is provided for a failed answer (Kornell, Hays, & Bjork, 2009; Pashler, Cepeda, Wixted, & Rohrer, 2005), perhaps because of test-potentiated learning (students learn more when study follows a test; e.g., Izawa, 1970). Regardless, from a practical standpoint, the fact that repeated multiple-choice quizzes (with feedback) can enhance performance on novel exam items increases the applicability of using testing (quizzing) in the classroom, especially in larger classes where the time needed to score recall tests may be prohibitive.

Another aspect of the multiple-choice quizzes is that they involved clicker responding, a response technology that may itself foster learning. For instance, clickers allow teachers to immediately display the classroom response profiles to each question, which is potentially an additional motivational component. This component was not operative in the present experiments, as the responses from the class were not displayed to students. However, responding via clickers, rather than paper and pencil, may hold more interest for students. In a recent quasi-experimental study, Mayer et al. (2009) examined in-class multiple-choice quizzing with and without clickers in a college classroom. In one class, no quizzes were administered; in another, two to four clicker quiz questions were administered; and in a third, the same quiz questions were administered in paper-and-pencil format. Only the clicker-administered quizzes produced significant gains in exam performance on different questions that targeted similar general content. Interpreting this finding as a benefit of clicker responding *per se* is not straightforward, though, because the clicker questions were integrated into classroom presentations followed by discussion of how to answer the quiz items, whereas the paper-and-pencil questions were administered at the end of the class and with no discussion. With regard to the present findings, perhaps the clickers introduced an element of fun for the students that contributed to learning. In initial clicker quiz experiments in social studies at the middle school that participated in the current study, Roediger, Agarwal, et al. (2011) reported that the students indicated that the use of clickers in class was enjoyable.

Further possible interpretations of the present quizzing results rest on the observation that classroom studies lack control over students' behavior during the interval between quizzes and the exam. As a consequence, indirect benefits of testing (e.g., Roediger, Putnam, & Smith, 2011) may also have contributed to the middle-school students' improved performance on previously quizzed content. For example, laboratory studies suggest that testing helps students identify gaps in knowledge (Son & Kornell, 2008) and improves their evaluation of future performance (Kornell & Son, 2009; Thomas & McDaniel, 2007), which could inform what concepts students choose to study during the interval between quizzes and the exam. Thus, in addition to (or instead of) any direct benefits of spaced quizzing with feedback for retention and near transfer, quizzing in the present study may also have improved metacognitive evaluation of learning and increased the effectiveness of subsequent study behavior.

As we have repeatedly emphasized, another limitation of the present experiments is that there was no condition that controlled for re-exposure to the quizzed items. As a result, the exam questions for quizzed material may have benefitted relative to the exam questions that were not quizzed because of additional exposure of the content (which would include exposure through the questions stems and answers, as well as feedback provided to the quizzed items), because of the spaced nature of the additional exposure, because sometimes the quizzed items were presented more recently to the exam than nonquizzed material, or some combination of these factors. Further, students may have assumed that quizzed material was more important than nonquizzed material and devoted more study time to that material when preparing

for the exams. However, experiments that have included a restudy control condition along with the a quizzing condition have usually shown that quizzing is superior to restudying, both in the laboratory (Roediger & Karpicke, 2006b) and in middle-school classrooms such as those participating in the present study (Roediger, Agarwal, et al., 2011). Also, term-response quiz items did not increase performance on application exam items relative to the no-quiz control items (Experiment 2). This finding suggests that the quiz effect was not necessarily a consequence of students' devoting additional study to quizzed information in preparation for the exam. Still, the possibility remains that the quizzes produced positive effects because of additional study or exposure to the target content relative to that for the no-quiz items (see e.g., Kang et al., 2007; McDaniel et al., 2007, for findings along these lines when multiple-choice quizzes are used).

In sum, this experimental study establishes that a spaced quizzing procedure with feedback can enhance performance on classroom exams that contain novel questions on quizzed content. This transfer was relatively broad, ranging from increased learning of definitional information (after various kinds of questions including term-response and applied questions) to more accurate application of scientific principles. Thus, quizzing (testing) implemented with components such as repetition, spacing, and feedback appears to be a technique that enhances learning of science concepts, not just learning of particular answers to repeated questions (cf. McDaniel et al., 2011; Roediger, Agarwal, et al., 2011). From a practical perspective, regardless of the possible contributions of the factors just discussed to the present quizzing effects (factors that were not linked to quizzing *per se*), quizzing is at the least a technique to effectively repeat and space material in a fashion that actively engages students (requires production of responses). The upshot is that there is pedagogical value to implementing quizzes relative to not doing so for enhancing performance on summative tests. An advantage of quizzing (testing) is that it can be incorporated in a wide variety of educational contexts, without extensive changes or adjustments to current classroom practice and teacher development.

ACKNOWLEDGEMENTS

This research was supported by grant R305H060080-06 to Washington University in St. Louis from the Institute of Education Sciences, US Department of Education. The opinions expressed are those of the authors and do not represent the views of the Institute or the US Department of Education. We are grateful to the Columbia Community Unit School District 4; superintendents Leo Sherman, Jack Turner, and Ed Settles; Columbia Middle School principal Roger Chamberlain; science teachers Teresa Fehrenz and Ammie Koch; and the seventh-grade and eighth-grade students and their parents. We also thank Lindsay Brockmeier, Lisa Cressey, and Barbie Huelser for their help in preparing materials and testing students and Jane McConnell, Kari Farmer, and Jeff Foster for their assistance throughout the project.

REFERENCES

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.

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

Barnett, S.M., & Ceci, S.J (2002). When and where to we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, *128*, 612–637.

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

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

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

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

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review*, *13*, 826–830.

Chan, C. K., McDermott, K. B., & Roediger, H. L. (2006). Retrieval induced facilitation: Initially nontested material can benefit from prior testing. *Journal of Experimental Psychology. General*, *135*, 533–571.

Daniel, D. B., & Broida, J. (2004). Using Web-based quizzing to improve exam performance: Lessons learned. *Teaching of Psychology*, *31*, 207–208.

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

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

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

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

Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology*, *101*, 621–629.

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

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *15*, 966–968.

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.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction integration model. *Psychological Review*, *95*, 163–182.

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

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.

Larsen, D. P., Butler, A. C., & Roediger, H. L. (2008). Test-enhanced learning in medical education. *Medical Education*, *42*, 959–966.

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

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

Lyle, K. B., & Crawford, N. A. (2011). Retrieving essential material at the end of the lecture improves performance on statistics exams. *Teaching of Psychology*, *38*, 94–97.

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.

McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology*, *16*, 192–201.
McDaniel, M. A., & Masson, M. E. J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 371–385.
McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*.
McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*, 494–513.
McDaniel, M. A., Friedman, A., & Bourne, L. E. (1978). Remembering the levels of information in words. *Memory & Cognition*, *6*, 156–164.
McDaniel, M. A., Howard, D. C., & Einstein, G. O. (2009). The read–recite–review study strategy: Effective and portable. *Psychological Science*, *20*, 516–522.
McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative-assessment performance in a Web-based class: An experimental study. *Journal of Applied Research in Memory and Cognition*, *1*, 18–26.
Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer-appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 519–533.
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.
Popham, W. J. (2011). *Classroom assessment: What teachers need to know* (6th edn.). Boston, MA: Allyn & Bacon.
Pyc, M.A., & Rawson, K.A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory?. *Journal of Memory and Language*, *60*, 437–447.
Roediger, H. L., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210.
Roediger, H. L., & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*, 249–255.
Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology. Applied*, *17*, 382–395.
Roediger, H. L., Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. In J. Mester, & B. Ross (Eds.), *The psychology of learning and motivation: Cognition in education* (pp. 1–36). Oxford: Elsevier.
Roediger, H. L., Weldon, M. S., & Challis, B. H. (1989). Explaining dissociations between implicit and explicit measures of retention: A processing account. In H. L. Roediger, & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honor of Endel Tulving* (pp. 3–41). Hillsdale, NJ: Erlbaum.
Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 233–239.
Son, L. K., & Kornell, N. (2008). Research on the allocation of study time: Key studies from 1890 to the present (and beyond). In J. Dunlosky, & R. A. Bjork (Eds.), *A handbook of memory and metamemory* (pp. 333–351). Hillsdale, NJ: Psychology Press.
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*]. Texas: eInstruction Corporation.
Zaromb, F. M., & Roediger, H. L. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & Cognition*, *38*, 995–10.

APPENDIX
QUIZ AND EXAM QUESTION EXAMPLES

	Experiment 1	
	Term response	Definition response
Quizzes: cells	What process is used when a cell needs to take in a substance that is higher in concentration inside the cell then outside and requires the cell to use energy to complete this process?	What is active transport?
	A. Passive transport	A. When a cell moves water without the use of energy.
	B. Active transport	B. The movement of RNA from the Golgi body to the nucleus.
	C. Osmosis	C. The transportation of DNA from the endoplasmic reticulum to the nucleus.
	D. Diffusion	D. The movement of material through the cell membrane using energy.
Exam: cells	What process is the movement of materials through a cell membrane using energy?	Which of the following correctly describes active transport?
Quizzes: machines/energy	What is the measure of the average kinetic energy of the individual particles in an object?	What is the definition of temperature?
	A. Heat	A. A thermal energy scale that measures heat that is transferred from one substance to another.
	B. Thermal energy	B. The measure of the average kinetic energy of the particles in a substance.
	C. Temperature	C. A scale that measures the net amount of kinetic energy of particles in a substance.
	D. Energy	D. A scale that measures the amount of particles in a substance.
Exam: machines/energy	What is the average kinetic energy of an object's individual particles?	How would you define temperature?
Quizzes: animals	What is the process in which newly hatched birds or newborn mammals learn to follow the first object they see?	What is imprinting?

(*Continues*)

(Continued)

	A. Imprinting	A. A process of performing complex tasks like learning and solving problems in newly hatched birds or newborn mammals.
	B. Cloning	B. A process in which newly hatched birds exhibit inborn behavior patterns that an animal performs correctly the first time.
	C. Adapting	C. A process in which newly hatched birds or newborn mammals learn to follow the first object they see.
	D. Instinct	D. A process of learning that requires continuous exposure to pictures that illustrate behavior in newly hatched birds or newborn mammals.
Exam: animals	When newly hatched birds or newborn mammals learn to follow the first object they see, what is this called?	Which of the following describes imprinting?
	Experiment 2	
	Term response	Application
Quizzes: ecology	What is the struggle between organisms to survive in a habitat with limited resources?	Both foxes and raccoons on Long Island eat pheasant, which in recent years, has been in decline. The foxes and raccoons' situation is an example of what ecological process?
	A. Parasitism	A. Parasitism
	B. Limited factors	B. Competition
	C. Predation	C. Limiting factors
	D. Competition	D. Predation
Exam: ecology	What is the term for when two or more organisms vie for limited environmental resources?	A group of 500 pandas are living in a reserve. Recent dry weather has reduced the bamboo populations, which the pandas rely on. The pandas are in what type of relationship?
Quizzes: environment	Which of the following represents the idea that humans have the right and ability to use resources from the Earth without restraint, especially those that will benefit humans?	A person who says, 'I will use all the coal available if that is what is economical, even if it hurts the environment', would hold which of the following viewpoints?
	A. Preservation viewpoint	A. Preservation viewpoint
	B. Conservation viewpoint	B. Manifest viewpoint
	C. Development viewpoint	C. Conservation viewpoint
	D. Manifest viewpoint	D. Development viewpoint
Exam: environment	Using Earth's resources without restraint when it benefits humans characterizes which of the following viewpoints?	'Despite the damage caused to forests, I will continue to use as much paper as I want. Since it is the most convenient for me, it does not matter how it might affect the environment.' Someone who believes this holds which viewpoint?

Note. For Experiment 2, the multiple-choice quiz and exam questions had identical answer options for questions of the same type, although the order of the answer options varied across quizzes and the exam. The correct answers for each question are underlined