Abstract Title Page

Not included in page count.

Title: Quizzing Promotes Deeper Acquisition in Middle School Science: Transfer of Quizzed Content to Summative Exams

Author(s):

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

Abstract Body

Limit 5 pages single spaced.

Background / Context:

Description of prior research and its intellectual context.

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 basic literature has established that testing is not a neutral event—testing can also enhance and modify memory (Carpenter & DeLosh, 2006; Hogan & Kintsch, 1971; McDaniel & Masson, 1985; see Roediger & Karpicke, 2006, for a review). Such findings suggest that educators might exploit testing (e.g., no- or low-stakes quizzing) 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 no- and low-stakes 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., Leeming, 2002; see Bangert-Drowns, Kulik, & Kulik, 1991, for a review). These patterns have been reinforced by recent experimental studies in college courses (McDaniel, Wildman, & Anderson, 2010) and middle school courses (McDaniel, Agarwal, Huelser, McDermott, & Roediger, in press; Roediger, McDaniel, McDermott, & Agarwal, 2010) showing significant improvement on course summative assessments for material that has previously appeared on no- or low-stakes 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, however, to nearly all of the laboratory and the classroom experimental demonstrations of the testing effect is that the summative assessment (final test) questions have been the same as those used for the quizzes (e.g., Carpenter, Pashler, & Cepeda, 2009; McDaniel et al., in press; Roediger et al., 2010). In some educational contexts, providing identical items on the quiz and the summative assessment 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 course contexts (McDaniel et al., in press). This context notwithstanding, many educators and educational theorists would strongly object to including summative test items on initial quizzes (Popham, 2011). Accordingly, most extant experimental studies of the testing effect do not necessarily compel its broad utility.

Yet, there are theoretical reasons and associated recent laboratory work that favor the idea that testing would benefit performance on summative assessment items that are related but not identical to the items presented on the initial test (quiz). First, testing improves associative learning and retention relative to additional study of material (e.g., for learning the meaning associated with a new vocabulary item, see Karpicke & Roediger, 2008; for word pair materials, see Carpenter, Pashler, & Vul, 2006). To the extent that the acquired associations are bi-directional (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. Support for this expectation was recently reported in two laboratory experiments with fourth and fifth graders learning to associate county or city names with locations on fictional maps (Rohrer, Taylor, & Sholar, 2010). 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 previously associated element as the answer ("All _______ axons, whether sympathetic or parasympathetic, release acetylcholine as a neurotransmitter"; McDaniel, Anderson, Derbish, & Morrisette, 2007).

Second, recent laboratory experiments hint that testing might also stimulate deeper learning. Chan, McDermott, and Roediger (2006) found that testing produced better performance on related but untested information, suggesting that testing may produce more extensive activation of information related to the question that is not required for the answer itself. In Butler (2010), 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 it received higher scores on inference and applied questions compared to subjects who reread the passage without intervening recall. These findings imply that testing can produce more complete acquisition of constructs, perhaps including a more organized (Zaromb & Roediger, in press) or detailed mental model of the target information.

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

In light of the suggestive laboratory findings just reviewed, we thought it possible that low-stakes quizzing in the classroom might also prompt deeper or more complete learning of the course material, such that performance on course summative assessment items that required transfer of the tested information would be enhanced relative to no quizzing. To examine this possibility, we conducted three experiments in an authentic classroom situation in which performance on the summative examinations used to evaluate the students (and assign grades) served as our dependent measures. Of interest was the extent to which in-class quizzes (with feedback) would enhance performance on summative exams, especially when quiz items are related to, but are not identical to, items on the summative exam. Echoing the variety of transfer effects produced by testing (quizzing) reported across the tantalizing but limited available experimental work (see Rohrer et al., 2010), the present study was designed to explore a range of possible transfer from quizzed items to exam items. As an overview, Experiment 1 focused on the extent to which quizzing would produce associative transfer, and Experiments 2a and 2b examined the effects of quizzing on learning and retention of related information and application of target constructs.

Setting:

Description of the research location.

Students in Columbia Middle School (CMS) in Illinois served as participants. The school is located in Columbia, Illinois, a community about 25 minutes southeast of St. Louis. The research team has met many times with teachers, administrators of the school (Principal, Assistant Principal), and administrators of the School District (Curriculum Coordinator, District Superintendent). CMS enrolls students in grades 5-8, with a total enrollment of about 530 students. During the past three years, we have created a positive, enthusiastic, and cooperative atmosphere with CMS students, teachers, administrators, and parents.

Population / Participants / Subjects:

Description of the participants in the study: who, how many, key features or characteristics.

Approximately 150 7th students and 150 8th grade middle school students, including special education and gifted students, participated in this research. Students at CMS are about half male and half female. Ninety-seven percent of students are Caucasian. The principal of the nearby high school (in the same school district) estimates that 75% of the graduating seniors go on to some form of further education (including community colleges and technical trade schools).

Intervention / Program / Practice:

Description of the intervention, program or practice, including details of administration and duration.

In Experiment 1, the type of question changed from quiz to summative exam, such that the associative order of summative assessment items was the reverse of that presented in the quizzed items. Thus, superficial learning of a particular response from practice on quizzes would not be sufficient to support performance on these criterial questions. Specifically, for exam items that provided the concept term in the stem and required a definition for the response (for ease of exposition, we term these *definition* questions), the quiz items provided the definition in the stem and required the concept term for the response (we term these *concept-term* questions; see Appendix B, Table 1 for examples of questions). Items were quizzed in the same format for pre-lesson, post-lesson, and review quizzes, and the phrasing of the question remained the same for all three initial quizzes. Importantly, question stems on the unit exam were reworded so that none of the questions from the initial quizzes was identical to the unit exam questions.

In Experiments 2a and 2b, we examined how quizzing might impact performance on exam items that required students to figure out what principle or construct was being illustrated in a particular scenario or situation (we label these *application* questions; see Appendix B, Table 1). In both experiments, one-third of items were initially quizzed (on pre-lesson, post-lesson, and review quizzes) in a concept-term format, one-third were initially quizzed in an application format, and one-third were not quizzed. At the end of the unit, students received concept-term questions on half of the items and application questions on half of the items, such that items were in each of the six conditions generated by the factorial combination of quiz format (concept-term format, application format, no quiz) and unit exam format (concept-term, application). Each of the six classroom sections had a different random assignment of items to the six conditions.

Research Design:

Description of research design (e.g., qualitative case study, quasi-experimental design, secondary analysis, analytic essay, randomized field trial).

We used a true experimental design, in which the manipulated intervention occurred within-student, such that some materials received normal classroom exposure and other materials were assigned to the treatment condition (additional quizzing), with materials counterbalanced across students. This within-students design feature provides several advantages to the more common between-classroom, between–students design. First, power is maximized. The classroom experiments conducted in our project had extremely high power to detect a .10 effect (a small size effect): power = .995 (alpha = .05, two-tailed). Second, the within-students design precludes the potential ethical issue associated with designs in which some students have potential benefits in course performance (because of the testing intervention) and other students shoulder the costs of being deprived of the testing intervention (no-test control). Indeed, the Columbia school administrators raised this concern during our initial contacts with them, stimulating our implementation of within-subject manipulations.

Data Collection and Analysis:

Description of the methods for collecting and analyzing data.

To measure retention, the classroom teacher administered unit exams in paper-and-pencil format. Students completed a multiple-choice test comprised of all quizzed and non-quizzed items. Initial quiz and final unit exam performance was analyzed using repeated measures of analysis of variance (ANOVA).

Findings / Results:

Description of the main findings with specific details.

Across all three experiments (see Appendix B, Table 2), initial quiz performance increased from pre-lesson to post-lesson and review quizzes, regardless of question type. Regarding unit exam performance for Experiment 1 (with 7th grade students; see Appendix B, Figure 1), there was a significant main effect of quiz question type, F (2, 120) = 82.97, MS_e = .010, η_p^2 = .58, such that exam performance was enhanced when the target content had been previously quizzed relative to unquizzed content. Further, a significant interaction between quiz question type and exam question type, F (2, 120) = 5.93, MS_e = .009, η_p^2 = .09, indicated that the benefits of quizzing (relative to no quizzing) were slightly more pronounced when the quiz question was the same type as the exam question (e.g., concept-term quiz—concept-term exam question) compared to when the quiz question was a different type (e.g., definition quiz—concept-term exam question). To directly evaluate whether quizzing improved exam performance for both same-type exam questions and different-type exam questions (relative to no quizzing), we conducted two sets of planned comparisons. The first set showed that students scored higher on concept-term exam questions after being quizzed with corresponding concept-term questions compared to not quizzed concept-term questions, F (1, 120) = 35.26, MS_e = .009, d = 0.81. Similarly, students scored higher on definition exam questions after being quizzed with corresponding definition questions compared to not quizzed definition questions, F (1, 120) = 118.51, MS_e = .009, d = 1.33. Next, we considered whether test-enhanced learning occurred even when the stem of the question changed from initial quizzes to the unit exam. Students scored higher on definition exam questions when they had been quizzed with concept-term questions compared to not quizzed definition questions, F (1, 120) = 87.84, MS_e = .009, d = 1.23. Likewise, students also scored higher on concept-term exam questions when they had been quizzed with definition questions compared to not quizzed concept-term questions, F (1, 120) = 38.08, MS_e = .009, d = 0.86.

Regarding unit exam performance for Experiment 2a (with 7th grade students; see Appendix B, Figure 2), there was a significant main effect of quiz question type, F (2, 188) = 10.29, MS_e = .028, η_p^2 = .10, such that exam performance was enhanced when the target content had been previously quizzed relative to unquizzed content. The interaction between quiz question type and exam question type was not significant (F < 1). Next, we directly tested whether quizzing improved exam performance for both same-type and different-type exam questions (relative to unquizzed). As in Experiment 1, students scored higher on concept-term exam questions after being quizzed with corresponding concept-term questions compared to not quizzed concept-term questions, F (1, 188) = 13.40, MS_e = .032, d = 0.62. Students tended to score higher on application exam questions after being quizzed with corresponding application questions compared to not quizzed application questions, F (1, 188) = 2.93, MS_e = .032, d = 0.20, p = .09. Next, we considered whether test-enhanced learning occurred even when the type of initial quiz question differed from the exam question. Students scored higher on application exam questions when they had been quizzed with concept-term questions compared to not quizzed questions, F (1, 188) = 4.01, MS_e = .032, d = 0.25. Likewise, students scored higher on concept-term exam questions when they had been quizzed with application questions compared to not quizzed questions, F (1, 188) = 8.13, MS_e = .032, d = 0.45.

Regarding unit exam performance for Experiment 2b (with 8th grade students; see Appendix B, Figure 3), there was a significant main effect of quiz question type, F (2, 178) = 9.95, MS_e = .028, η_p^2 = .10, such that exam performance was enhanced when the target content had been previously quizzed relative to unquizzed content. Further, a significant interaction between quiz question type and exam question type, F (2, 178) = 4.79, MS_e = .034, η_p^2 = .05, revealed that the benefits of quizzing were slightly more pronounced when the quiz question was the same type as the exam question (e.g., application quiz—application exam question) compared to when the quiz question was a different type (e.g., concept-term quiz—application exam question). The next set of analyses aimed to directly evaluate whether quizzing improved exam performance for both same-type and different-type exam questions (relative to unquizzed). Students scored higher on concept-term exam questions after being quizzed with corresponding concept-term questions compared to not quizzed concept-term questions, F (1, 178) = 21.35, MS_e = .034, d = 0.56. Likewise, students scored higher on application exam question after being quizzed with corresponding application questions compared to not quizzed application questions, F (1, 178) = 4.77, MS_e = .034, d = 0.27. Next, we considered whether test-enhanced learning occurred even when the question type changed focus from definitional to application (or vice versa) between the initial quiz and exam. Students scored higher on concept-term exam questions when they had been quizzed with application questions compared to not quizzed questions, F (1, 178) = 7.65, MS_e = .034, d = 0.34. However, students did not score any higher on application exam questions when they were quizzed with related concept-term questions compared to the not quizzed questions (F < 1).

Conclusions:

Description of conclusions, recommendations, and limitations based on findings.

The current study significantly extends, from both a theoretical and a practical perspective, related experiments on quizzing effects for improving summative assessment performances on which students' grades are based in authentic classrooms. In previous experiments conducted in middle school classes, the items on the exams were identical to those presented on the quizzes (McDaniel et al., in press; Roediger et al., 2010). In these cases, the benefits for the exam performances could rest simply on retrieval practice of particular answers during quizzing. The present results demonstrate that low- and no-stakes quizzing can promote learning that is deeper than just retaining a particular answer. Experiment 1 clearly showed that quizzing promoted transfer to different exam items requiring a reverse association between concept-term and definition from that quizzed. Experiments 2a and 2b further showed that quizzing promoted transfer from applying a principle/concept in a concrete context to better retention of definitional information, as well as to applying the principle in a new context. This transfer was relatively broad, ranging from associative transfer, to increased learning of definitional information (after applied questions), to application of concepts in a variety of situations. Thus, quizzing can enhance learning of science concepts, not just learning of particular answers to repeated questions (across quizzes and exams). As such, low- or no-stakes quizzing appears to be a valuable learning technique that could be incorporated in a wide variety of educational contexts, without extensive changes or adjustments to current classroom practice and teacher development.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

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. *Journal of Educational Research*, *85*, 89-99.

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.

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.

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. Homa, D. & Vosburgh, R. (1976). Category breadth and the abstraction of prototypical information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *2*, 322-330.

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.

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.

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (in press). 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., Howard, D. C., & Einstein, G. O. (2009). The read-recite-review study strategy: Effective and portable. *Psychological Science, 20*, 516-522.

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

McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2010). *Classroom testing effects with repeated quizzes*. Manuscript submitted for publication.

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210.

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

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

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

Appendix B. Tables and Figures
Not included in page count.

Table 1
Quiz and Unit Exam Question Examples

Experiment 1

	Concept-term	Definition
Quiz	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 <u>B. Active Transport</u> C. Osmosis D. Diffusion	What is 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. <u>D. The movement of material through the cell membrane using energy.</u>
Unit Exam	What process is the movement of materials through a cell membrane using energy?	Which of the following correctly describes active transport?

Experiment 2a

	Concept-term	Application
Quiz	What rule in physics states that a stream of fast moving fluid exerts less pressure than the surrounding fluid? A. Mead's Principle <u>B. Bernoulli's Principle</u> C. Piaget's Principle D. Erikson's Principle	When Sally is at home by the fireplace, smoke rises up the chimney because hot air rises, and partly because it is pushed by the wind blowing across the top of the chimney. This lowers the overall air pressure causing the high pressure at the bottom to push the smoke up. What principle keeps smoke from filling up the room? A. Mead's Principle <u>B. Bernoulli's Principle</u> C. Piaget's Principle D. Erikson's Principle

Unit Exam	What rule in physics states that as the velocity of a fluid increases, the pressure exerted by that fluid decreases?	When a pitcher throws a curve ball, the spin of the ball creates high pressure on top of the ball, which pulls the ball downward. What principle is being illustrated in this example?

Experiment 2b

	Concept-term	Application
Quiz	What is the struggle between organisms to survive in a habitat with limited resources? A. Parasitism B. Competition C. Limited Factors D. Predation	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 B. Competition C. Limiting Factors D. Predation
Unit Exam	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?

Note. The multiple-choice quiz and unit exam questions had identical answer options for questions of the same type, although the order of the answer options varied.

Table 2
Initial quiz performance (proportion correct) as a function of quiz placement and initial quiz question type for all three experiments.

	Initial Quiz Placement		
	Pre-lesson	Post-lesson	Review
Experiment 1 (n = 61)			
Concept-term	.48 (.01)	.73 (.01)	.82 (.01)
Definition	.49 (.01)	.69 (.02)	.78 (.02)
Experiment 2a (n = 95)			
Concept-term	.52 (.02)	.79 (.02)	.90 (.01)
Application	.54 (.02)	.77 (.02)	.89 (.01)
Experiment 2b (n = 90)			
Concept-term	.56 (.02)	.82 (.02)	.85 (.02)
Application	.51 (.02)	.79 (.02)	.86 (.02)

Note. Standard error is noted in parentheses.

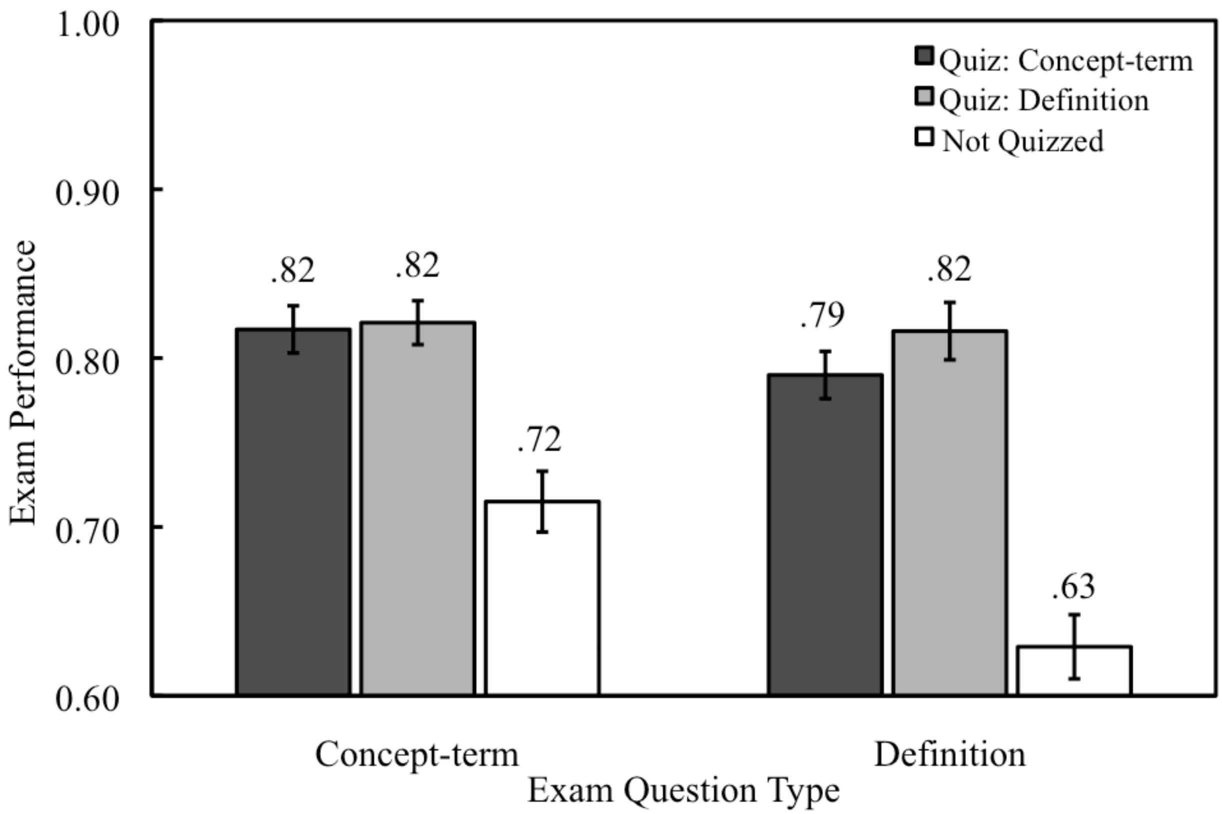

Figure 1. Unit exam performance (proportion correct) on concept-term and definition questions as a function of initial quiz question type in Experiment 1. Error bars represent standard error of the mean.

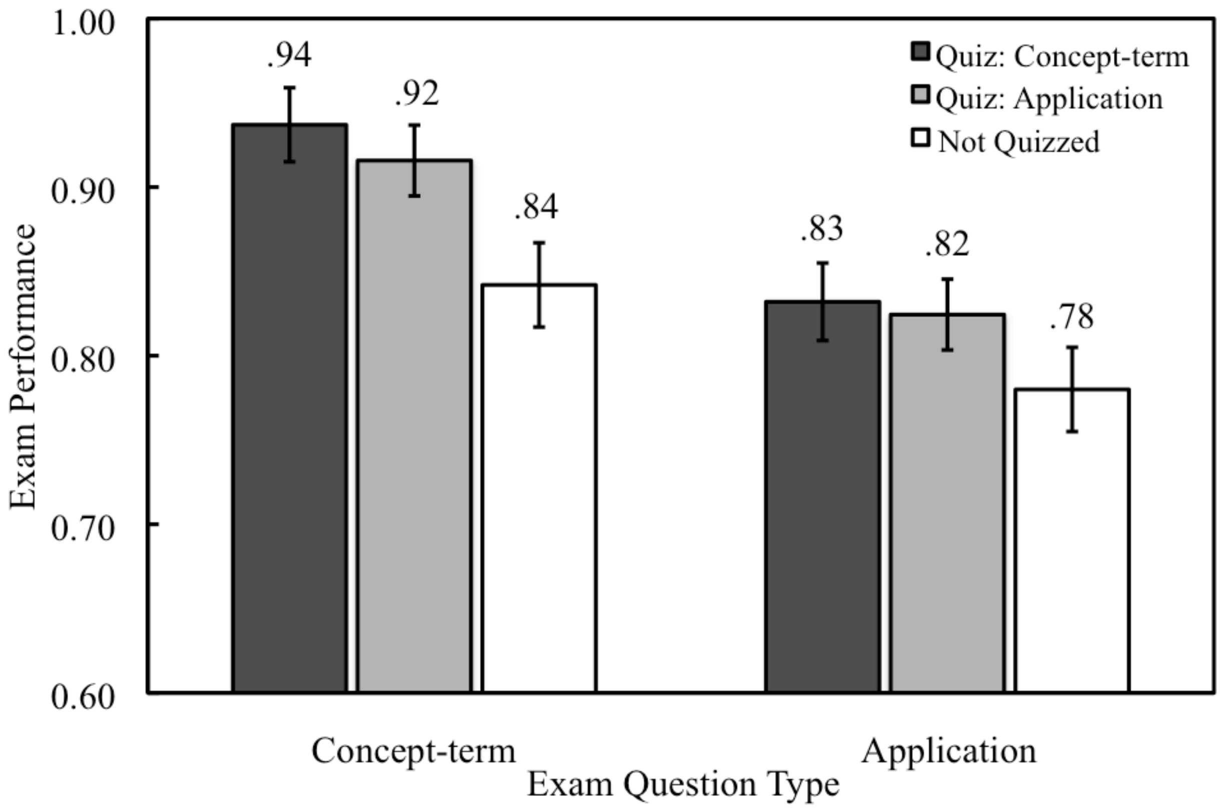

Figure 2. Unit exam performance (proportion correct) on concept-term and application questions as a function of initial quiz question type in Experiment 2a. Error bars represent standard error of the mean.

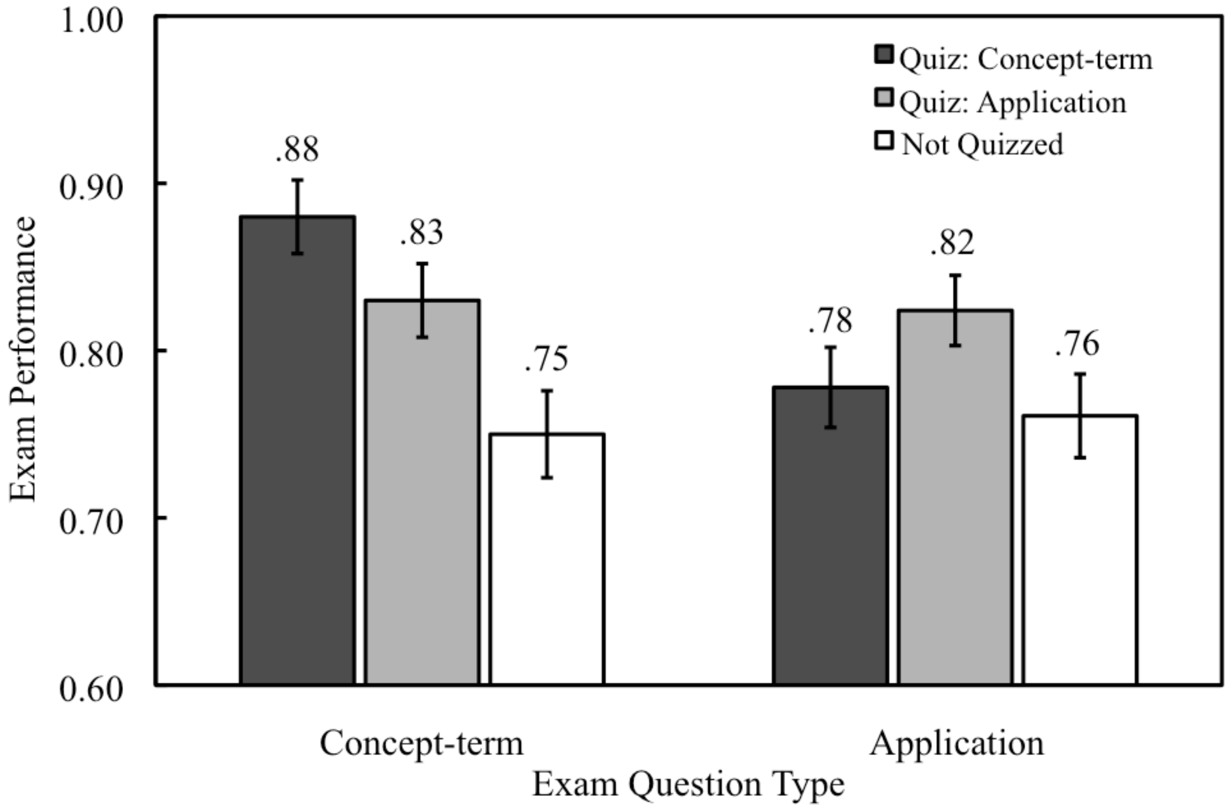

Figure 3. Unit exam performance (proportion correct) on concept-term and application questions as a function of initial quiz question type in Experiment 2b. Error bars represent standard error of the mean.