Effects of Reinforcement on Test-Enhanced Learning in a Large, Diverse Introductory College Psychology Course

Michael C. Trumbo and Kari A. Leiting
The University of New Mexico

Gordon K. Hodge
Washington University

A robust finding within laboratory research is that structuring information as a test confers benefit on long-term retention—referred to as the testing effect. Although well characterized in laboratory environments, the testing effect has been explored infrequently within ecologically valid contexts. We conducted a series of 3 experiments within a very large introductory college-level course. Experiment 1 examined the impact of required versus optional frequent low-stakes testing (quizzes) on student grades, revealing students were much more likely to take advantage of quizzing if it was a required course component. Experiment 2 implemented a method of evaluating pedagogical intervention within a single course (thereby controlling for instructor bias and student self-selection), which revealed a testing effect. Experiment 3 ruled out additional exposure to information as an explanation for the findings of Experiment 2 and suggested that students at the college level, enrolled in very large sections, accept frequent quizzing well.

Keywords: testing effect, retrieval practice, classroom learning, education

Traditionally, lecture, discussion, and study are geared toward acquisition of knowledge, and testing is viewed as a way to assess the degree of mastery over such knowledge (Graue, 1993). Tulving (1967), however, demonstrated that tests themselves might be conducive to learning, rather than simply serving as neutral assessment events. Shortly thereafter, additional research demonstrated that incorporating a testing event during a learning phase is likely to enhance recall following a delay, relative to a learning phase that includes study events but no testing (Hogan & Kintsch, 1971). Since Tulving’s work, additional studies have supported the notion that testing enhances long-term retention, relative to the benefit of repeated study (e.g., Hogan & Kintsch, 1971; Karpicke & Roediger, 2007; Thompson, Wenger, & Bartling, 1978; Wheeler, Ewers, & Buonomano, 2003; see Roediger & Karpicke, 2006a for a review). Despite robust and replicated findings, research in this domain has largely failed to translate into practice (Glover, 1989; Pashler, Rohrer, Cepeda, & Carpenter, 2007; Roediger & Karpicke, 2006a).

Translating findings from laboratory studies to classroom practice may be hindered by the issue of ecological validity. Laboratory settings control exposure to information, often limiting exposure to a single session (e.g., Glover, 1989; Kang, McDermott, & Roediger, 2007), while classroom learning often involves repeated, varied, and spaced presentation of integrated content through multiple methods (e.g., lecture, homework, reading). Additionally, laboratory research often utilizes identical questions for initial and final tests (e.g., Kang et al., 2007; Roediger & Marsh, 2005), but course instructors may wish to vary the questions used on quizzes and tests (as we did in the current Experiments 2 and 3). A recent survey of nearly 200 introductory psychology teachers indicated that identical quiz and test questions are highly atypical in a college classroom environment (Wooldridge, Bugg, McDaniel, & Liu, 2014).

Despite possible lack of ecological validity, early testing effect research within classroom settings has been encouraging (Carpenter, Pashler, & Cepeda, 2009; Duchastel & Nungester, 1982; Sones & Stroud, 1940; Swenson & Kulhavy, 1974) and has produced a nascent body of research demonstrating that the testing effect may occur under real-world circumstances. Research using actual
grated course assessments has revealed testing effects in college courses (e.g., Bjork, Little, & Storm, 2014; McDaniel, Wildman, & Anderson, 2012) as well as middle and high school classrooms (McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011; McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013; McDermott, Agarwal, D’Antonio, Roediger, & McDaniel, 2014; Roediger, Agarwal, McDaniel, & McDermott, 2011).

Under some circumstances, however, quizzing does not seem to benefit learning. In contrast to the findings of McDaniel and colleagues (2012), Daniel and Broida (2004) found online quizzes ineffective at the college level relative to in-class quizzes, unless certain conditions were imposed (e.g., random assignment of questions from a larger test bank, limited time per quiz attempt). Bell, Simone, and Whitfield (2015) found that the use of online quizzing failed to improve in-class quiz and exam scores for introductory psychology students.

It is likely that the context and the particulars of quiz implementation matter. While findings for upper level college courses (e.g., research methods, Bjork et al., 2014; child and adolescent development, Daniel & Broida, 2004; psychology statistics, Lyle & Crawford, 2011) have demonstrated quizzing benefit, students in upper level courses are more likely to exhibit mastery goals which are predictive of course success (Harackiewicz, Barron, Tauer, & Elliot, 2002). To date, only one study (McDaniel et al., 2012) has found a benefit of implementing Web-based frequent low-stakes testing as a graded course component in an introductory level college course under controlled conditions (i.e., benefit of quizzing over restudy; see Glass, Brill, & Ingate, 2008 and Glass, 2009 for benefit of online and in class quizzes in a large introductory psychology course but without a control condition). It is worth noting that this benefit occurred in the context of a small (N = 16 in Experiment 1 and N = 27 in Experiment 2), entirely online course (McDaniel et al., 2012), making it uncertain if the findings would extend to an introductory level face-to-face course populated with hundreds of students as examined in the current work.

Although online quizzes provide the opportunity to incorporate frequent low-stakes testing into a course without sacrificing class time (Brothen & Wambach, 2001), teachers may shy away from the time and monetary investment required to adopt experimentally devised testing approaches to their real-life classrooms unless there is a clear indication that there is a benefit to doing so. Demonstration of a quizzing benefit in a large introductory college course that is both demographically and motivationally diverse is a critical step in determining generalizability of findings.

With these considerations in mind and to extend the practicality of online testing, the current work utilized required web-based, low-stakes, randomly generated, multiple-choice quizzes that provided correct-answer feedback following completion of each quiz attempt in the context of a large face-to-face introductory college-level course. Students were allowed an unlimited number of attempts for each set of quizzes leading up to a chapter test to take advantage of the benefits of frequent testing and overlearning (i.e., continuing to review information following successful recall), which has been found to increase the probability of future successful recall (Driskell, Willis, & Copper, 1992; Foriska, 1993).

There are a number of ways students may engage in retrieval practice, including flashcards, practice problems in textbooks, and classroom participation questions. These procedures allow students the opportunity to assess their degree of mastery over material, and are commonly used in this self-testing capacity (Hartwig & Dunlosky, 2012; Kornell & Bjork, 2007; Yan, Thai, & Bjork, 2014). However, college students list repeated reading of material as their most frequent study method and perceive restudying information to be the most effective study strategy (Karpicke, Butler, & Roediger, 2009; Yan et al., 2014). This is consistent with the predictions that students make during the course of laboratory research involving retrieval practice, in that students typically predict that restudy of information will maximize long-term retention, whereas retrieval practice is thought to be the weakest strategy (Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008; Karpicke & Blunt, 2011; Karpicke & Roediger, 2007). Ideally, students would voluntarily engage in optimal study strategies, but students do not appear to exhibit awareness of the mnemonic benefits of testing as a method of learning material, and often use testing instead purely as an assessment device.

Previous research comparing between course sections has demonstrated a benefit of implementing required quizzes relative to a section for which quizzes were not available (e.g., Daniel & Broida, 2004) and relative to restudy (e.g., McDermott et al., 2014). To our knowledge, no one has examined the extent to which students will engage in quizzing as a study strategy if quizzes are made available but not required as part of a course grade. When students learn material, even with the expectation of being tested on it, many students may require extrinsic motivation to invest a sufficient amount of time studying (Benware & Deci, 1984), particularly in an introductory-level course in which intrinsic motivation to master material may be lacking (Harackiewicz et al., 2002).

In Experiment 1, we provided a study strategy (quizzes) for students while we concomitantly manipulated extrinsic motivation to take quizzes. We required quizzes as part of the course grade in one section and made quizzes an optional study aid in another section. Extrinsic reinforcement in a classroom environment is most effective when presented on multiple occasions and when it is contingent on quality of performance rather than mere participation (Akin-Little, Eckert, Lovett, & Little, 2004). To that end, we provided extrinsic reinforcement in the quizzes-required section by giving a performance-based number of points for each quiz completion and allowing an unlimited number of quiz attempts. Our hypothesis was requiring quizzes in this fashion would motivate students to take them more often and obtain higher scores, leading to greater mastery of material as measured by test performance.

### Experiment 1: Required Versus Optional Quizzing

#### Method

**Participants.** Participants were 722 male and female students enrolled in two daytime sections of Introduction to Psychology at the University of New Mexico (UNM). They received course credit for their participation. While ethnicity data were not collected for this course, the ethnic composition of undergraduate students at UNM at the time of data collection was 50.5% Caucasian, 33.1% Hispanic, 6.4% American Indian, 3.4% Asian American, 2.7% African American, and 3.9% other or no response; approximately 58% of undergraduates at UNM identified as female (Office of Institutional Analytics [OIA], 2005).
Design. The experimental variable was whether quizzes were required or optional. The term quiz throughout this article is used to describe preparatory low-stakes online tests that students completed between exams. Material covering the entire textbook’s 17 chapters was divided into 15 units. For each of these units, Mastery Quiz A covered the first section of the unit and Mastery Quiz B covered the second. In addition, each unit was covered by a review quiz. Students in Section 001 (n = 427) were required to take weekly quizzes that covered material from each unit for course credit (labeled the quizzes-required section). Students in Section 002 (n = 295) were not required to take weekly quizzes; however, they were made available if students wanted to take them to study for the tests (labeled quizzes-optional section). The quizzes for both sections consisted of the same questions, and both sections had the same lectures and instructor. Students in both sections were told of the value of taking quizzes and were encouraged to take quizzes multiple times.

There were four tests administered over the course of the semester. The first three tests emphasized material in the three sections of the course; the fourth test was a cumulative final. We dropped students’ lowest test scores, including the final. We randomized quiz questions and alternative answers for all experiments. Exam questions were identical to quiz questions. We used four versions of tests, which we distributed so that no two students sitting next to each other received the same version. We proctored and checked student IDs.

Procedure. This research was conducted in compliance with the university’s Internal Review Board. Students received online access to a quiz during the week each chapter was being covered. They logged onto the course at their discretion to take the quiz. We instructed students that they could use the quizzes to study for tests. Additionally, we informed students not to refer to their textbook or lecture notes while taking quizzes, as this constitutes academic dishonesty (cheating).

WebCT served as the platform for online quiz administration. The quizzes for each unit became available by the day of the first lecture in the corresponding unit. Quiz questions were available all at once, allowing for backtracking. Students were given 15 min to complete each quiz. They were not required to respond to each question. Correct answer feedback for all questions appeared after the student quit or finished the quiz. There was no limit on the number of times students could take each quiz. Students had approximately one week to complete each successive quiz set; the last set was due before the test covering those quizzes. In the quizzes-required section, quiz scores counted for 45% of the students’ final grades. The students’ best scores on each of the three quizzes were summed, with a possible 88 points per unit (1 point possible per question, 20 questions for each of the two mastery quizzes; and 2 points possible per question, 24 questions for each review quiz).

We administered tests in class using paper and pencil. Each test consisted of 50 multiple-choice items. The scores on the four tests counted for a total of 50% of student final grade for those in the quizzes-required section and 94% of student final grade for those in the quizzes-optional section. Although we informed students of their grades on the tests, we allowed them to view their graded test by appointment only. To receive credit for the course, the university required a grade of C or higher; receiving a grade of C– or lower would result in the student not receiving credit for completing a university undergraduate core course requirement.

Results

Quiz performance. Students in the quizzes-required section took the quizzes more frequently (M = 4.19 SD = 2.35 vs. M = 0.65 SD = 1.04), t(720) = 24.21, p < .001, d = 1.83, 95% confidence interval [CI] [1.66, 2.01], obtained higher best scores on the quizzes (M = 14.92 SD = 4.73 vs. M = 2.60 SD = 3.43), t(720) = 35.84, p < .001, d = 2.72, 95% CI [2.5, 2.92], and spent more time on the quizzes (in minutes:seconds, M = 7:26 SD = 2:47 vs. M = 1:22 SD = 1:37), t(720) = 33.58, p < .001, d = 2.55, 95% CI [2.31, 2.73], than students in the quizzes-optional section. As shown in Table 1, this effect reflected consistent differences between sections within grade-outcome levels. Students who attained As in the quizzes-required section took more quizzes, spent more time, and achieved higher best scores than

<table>
<thead>
<tr>
<th>Table 1: Quiz Performance by Section and Course Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Course grade</strong></td>
</tr>
<tr>
<td>------------------</td>
</tr>
<tr>
<td><strong>Average number of attempts</strong></td>
</tr>
<tr>
<td>Required</td>
</tr>
<tr>
<td>Optional</td>
</tr>
<tr>
<td><strong>Average time (min)</strong></td>
</tr>
<tr>
<td>Required</td>
</tr>
<tr>
<td>Optional</td>
</tr>
<tr>
<td><strong>Average best score: mastery quizzes</strong></td>
</tr>
<tr>
<td>Required</td>
</tr>
<tr>
<td>Optional</td>
</tr>
<tr>
<td><strong>Average best score: review quiz</strong></td>
</tr>
<tr>
<td>Required</td>
</tr>
<tr>
<td>Optional</td>
</tr>
</tbody>
</table>

a Out of 20 points possible. b Out of 48 points possible.
* Indicates a significant difference (p < .001) between the value in the quizzes-required section and the value in the corresponding cell for the quizzes-optional section for each of the final course grades.
students who received As in the quizzes-optional section; this pattern was true for the other grades, as well. Differences (t tests) were significant overall, as well as between section grade levels, all ps < .001 (see Table 1).

**Test performance.** Pearson correlation was conducted to examine the relationship between total quiz points and total test points. For both sections, attaining higher best quiz points was correlated with higher test points, \( r = .76, p < .001; r = .87, p < .001 \), respectively. A t test on total exam scores showed that students in the quizzes-required section (\( M = 76.98\%, SD = 18.46\% \)) obtained higher scores on the tests, \( t(720) = 3.93, p < .001, d = 2.90, 95\% CI [2.68, 3.10] \), than students in the quizzes-optional section (\( M = 63.44\%, SD = 4.69\% \)).

Two one-way analyses of covariance were conducted to determine if there were statistically significant differences between sections on exam performance, one controlling for the number of quiz attempts and one controlling for average best quiz scores. Students in the quizzes-required section had higher scores on the tests than students in the quizzes-optional section even after controlling for number of quiz attempts, \( F(1, 719) = 375.64, p < .001, \eta^2_p = 0.34 \), and after controlling for average best quiz scores \( F(1, 719) = 717.04, p < .001, \eta^2_p = 0.49 \).

**Final grade performance.** There were significant differences in end of the semester course grades between the quizzes-required section and the quizzes-optional section, \( \chi^2(3, N = 722) = 44.19, p < .001 \). Follow-up chi-square tests were conducted and found that there were significantly more students in the quizzes-required section who achieved As as a final course grade (21%) than in the quizzes-optional section (8%), \( \chi^2(1, N = 120) = 5.34, p = .020, d = 0.43 \). Additionally, there were significantly more students in the quizzes-optional section who received a C– or below (53.9%) in the course compared to students in the quizzes-required section (32.9%), \( \chi^2(1, N = 299) = 5.141, p = .023, d = 0.26 \). A t test on final grade percentage confirmed that students in the quizzes-required section (\( M = 74.7\% \)) had higher course grades at the end of the semester than students in the quizzes-optional section (\( M = 66.7\% \)), \( t(720) = 5.32, p < .001, d = 0.40, 95\% CI [0.25, 0.55] \).

Finally, a one-way analysis of covariance was conducted to determine if there was still a statistically significant difference between sections on final course performance after controlling for average best quiz scores. Students in the quizzes-required section had better final grade performance than students in the quizzes-optional section even after controlling for average best quiz scores \( F(1, 719) = 36.48, p < .001, \eta^2_p = 0.048 \).

**Discussion**

Including required quizzes as part of the graded course assignments had a substantial impact on course outcome measures, as indicated by test scores and final grades. Most students who were not required to take quizzes were not apparently intrinsically motivated to take them and were not reinforced by completing them in the absence of performance-based course points, as shown by completion of fewer quizzes, less time spent on each quiz attempt, and lower best scores in the quiz-optional section (see Table 1). This pattern was consistent between sections within each grade-outcome level, demonstrating results were not biased by high performing students in the quizzes-required section or by low performing students in the quizzes-optional section. Students in the quizzes-required section had higher average best scores across all grade levels, typically with lower standard deviations than students in the quizzes-optional section. As there was a difference in test scores between sections after controlling for quiz attempts and best scores, a difference in final grades between sections remains even if the contribution of tests to final grades were to be equally weighted between sections. Students extrinsically motivated to take quizzes, as a group, consistently performed much better than students without extrinsic motivation. Results from our large and diverse population of students corroborates previous findings demonstrating students tend to score lower on tests when they do not take online quizzes (Kibble, 2007) or when they take fewer quizzes (Angus & Watson, 2009).

Presumably, students in both sections were motivated to obtain a high final grade in the course by performing well on the tests. In the absence of consistent performance-based reinforcement, students in the quizzes-optional section failed to utilize the quizzes as a study method to the same degree as their quiz-required counterparts. This demonstrates that students in an introductory psychology course may need extrinsic reinforcement to take advantage of quizzes, as they do not appear to find completion of quizzes intrinsically rewarding, despite the benefit quiz completion confers on test scores. These findings are consistent with survey research indicating that students typically engage in suboptimal study strategies such as highlighting (Bell & Limber, 2010) and rereading material (Hartwig & Dunlosky, 2012), so it may be necessary for instructors to require frequent testing as part of a course to take advantage of the testing effect. Likely labeling a method of learning as “optional” undermines the idea that such a method is a valuable learning tool. Additionally, students in the quiz-optional section may have viewed quizzes as an avoidable burden, and therefore failed to complete them to minimize the workload of the course.

The correlations between quizzes and tests illustrate how quiz performance predicts test performance and, by implication, course grades. Teachers and students are most concerned with grades. Structuring a course to include quizzes leads to higher scores across different course components and, eventually, higher grades. Teachers can use correlational information to highlight the benefit of quiz taking. Providing timely feedback of this sort to students may prove a more compelling source of motivation for students to take quizzes rather than engage in other study strategies. Even when controlling for number of quiz attempts and best quiz scores, students in the quizzes-required section performed better on tests than students in the quizzes-optional section. It is possible, therefore, that requiring quizzes has an indirect effect on test performance by encouraging diligent study habits (i.e., it may not be quizzes per se that improve performance). Additionally, although students chose a section in which to enroll, we did not inform them of course procedures beforehand; nevertheless, it is possible that a form of self-selection contributed to results. The instructor was the same for both sections, raising the possibility of instructor bias (although quizzes and tests were multiple-choice, allowing for objective grading, lack of instructor biasing potentially introduced differences in lecture delivery or student interaction). Although using a separate control section offers design advantages, insofar as required quizzing helped students we were reluctant to withhold quizzes from controls when testing additional
hypotheses. Further, exam questions were identical to quiz questions, limiting our ability to gauge conceptual learning.

We designed Experiment 2 with these concerns in mind. We split students within a single course section into groups that received selective exposure to particular information via an online delivery system. In this fashion, evaluation of a pedagogical manipulation controls for selection bias (via random group assignment within a single section) as well as instructor bias (all students receive the same lectures from the same instructor), and all students were able to benefit from the putative advantages of required quizzing. Also, exam questions were constructed so that they were not identical to quiz questions. We hypothesized that a testing effect would be evident when controlling for the influence of these factors, while allowing all students to benefit from the testing effect.

**Experiment 2: Evaluation of a Within-Course Manipulation**

**Method**

**Participants.** Participants were 715 male and female students enrolled in a daytime section of Introduction Psychology at UNM. They received course credit for their participation. The ethnic composition of undergraduate students at UNM at the time of data collection was 46.0% Caucasian, 35.1% Hispanic, 6.8% American Indian, 4.0% Asian American, 3.2% African American, and 4.9% other or no response; approximately 57% of UNM undergraduates identified as female (OIA, 2008).

As may be expected in any large college course, students joined and left the class over the semester, and sometimes failed to complete assignments; the number of students who dropped or exceeded the course was not tracked, and only students who completed the unit exam for a given unit were included for purposes of analyses. This resulted in 398 students to be included in the analysis.

We randomly assigned students to one of two groups to counterbalance quizzed and not quizzed content. The quizzed content for one group was not quizzed for the other group and vice versa. Given that students were randomly assigned to these counterbalancing groups, and that overall performances across these groups were statistically equivalent (see Results section), there is no reason to suspect that one group was differentially impacted by late adding/dropping students or students failing to complete assignments.

**Design.** All students were required to complete 3 quizzes (Mastery Quiz A, Mastery Quiz B, and Review Quiz). There were 11 unit exams, which we administered at the end of each unit, 6 days (on average) after lecture introduction of the material. Quizzes counted for 33% and unit exams for 50% of total points.

Across quizzes, wording of question stems remained the same but the order of the four alternative answers was randomized. The specific questions that appeared on each quiz attempt were randomly selected from larger subsets consisting of 98 possible questions (on average) for each mastery quiz (20 questions per quiz; 1 point per question) and 106 possible questions (on average) for each review quiz (16 questions per quiz; 2 points per question).

Of the possible questions, five critical questions (so called because related corresponding questions in the exams would probe for student memory) could appear on Mastery Quiz A and five others could appear on Mastery Quiz B. All 10 critical questions could appear on the review quiz, which consisted of questions from Mastery Quizzes A and B. The set of possible questions was divided into smaller groups based on topics within a given textbook chapter. The number of questions selected from each group depended on the perceived importance of the topic by the course instructor. Presentation of all possible critical quiz questions on each quiz attempt was not guaranteed due to the random selection of questions on each quiz attempt.

Each critical quiz question was paired with a corresponding exam question. We selected quiz questions from the first volume of the exam bank (Brink, 2008) that accompanied the required textbook (Myers, 2007). The relationships between the corresponding quiz and exam questions varied. For some corresponding questions, the stem of the quiz question became the answer to the exam question. In other cases, the quiz question directly tested knowledge of the fact and the exam question tested application of the fact. See Table 2 for example quiz and exam questions. At the end of each unit, students took the unit exam consisting of 50 multiple-choice questions with four alternatives (one correct answer, three lures) from the second volume of the exam bank (Brink, 2008). These questions consisted of 10 critical questions that were associated with the quiz-item sets and 10 critical questions that were not associated with the quiz-item sets. The 10 questions not associated with the quiz-item sets counted as extra credit. The remaining 30 questions on the unit exam were not associated with the quiz manipulation (all of these questions were potentially related to quizzed items); these 30 questions were drawn from items common to both groups.

**Procedure.** This research was conducted in compliance with the university’s Internal Review Board. We randomly assigned students to either of two conditions at the start of the semester.

<table>
<thead>
<tr>
<th>Question Type</th>
<th>Example Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quiz question</td>
<td>By directly experiencing a thunderstorm, we learn that a flash of lightning signals an impending crash of thunder. This best illustrates:</td>
</tr>
<tr>
<td></td>
<td>a. Operant conditioning</td>
</tr>
<tr>
<td></td>
<td>b. The law of effect</td>
</tr>
<tr>
<td></td>
<td>c. Observational learning</td>
</tr>
<tr>
<td></td>
<td>d. Classical conditioning</td>
</tr>
<tr>
<td>Exam question</td>
<td>Pet cats who learn that the sound of an electric can opener signals the arrival of their cat food illustrate:</td>
</tr>
<tr>
<td></td>
<td>a. Shaping</td>
</tr>
<tr>
<td></td>
<td>b. Extrinsic motivation</td>
</tr>
<tr>
<td></td>
<td>c. Classical conditioning</td>
</tr>
<tr>
<td></td>
<td>d. Observational learning</td>
</tr>
<tr>
<td>Brief statement</td>
<td>By directly experiencing a thunderstorm, we learn that a flash of lightning signals an impending crash of thunder. This best illustrates classical conditioning:</td>
</tr>
<tr>
<td></td>
<td>a. I’ll read this later</td>
</tr>
<tr>
<td></td>
<td>b. I’ve just read this but I know it now</td>
</tr>
<tr>
<td></td>
<td>c. I recognize this and I believe I’m learning it</td>
</tr>
<tr>
<td></td>
<td>d. I recognize this and I know it</td>
</tr>
</tbody>
</table>
Essentially, procedures for quizzes and exams were similar to those described in Experiment 1. After completing each set of three online unit quizzes, students in Experiment 2 completed an in-class, paper-and-pencil unit exam, for a total of 11 unit exams. Students were allowed unlimited quiz attempts and only their best score for a given quiz counted toward their grade.

Results

Quiz performance. A series of $t$ tests indicated that there were no significant differences across the two counterbalancing groups in terms of the number of times students took quizzes ($M = 6.37$, $SE = 6.85$), their best scores on quizzes ($M = 15.83$, $M = 15.43$), and time spent on quizzes ($M = 7.15$, $M = 6.58$), all $t$s < 1.0 and all $ps > .50$.

Unit exam performance. A repeated-measures analysis of variance showed that questions targeted by quizzes ($M = 7.08$, $SE = .029$) were answered correctly significantly more often than questions not targeted by quizzes ($M = 5.46$, $SE = .029$), $F(1, 199) = 2.70, p < .001, n^2_g = .07$. Figure 1 shows the mean differences between exam questions corresponding to the quiz sets and questions not corresponding to the quiz sets for all 11 units; the mean differences were significant ($ps < .001$) for all exams.

Additionally, there was no significant difference between the counterbalancing groups ($M = 20.15$, $M = 19.93$) on performance on the common questions, $r(198) = 1.54, p = .125$, on exam performance ($M = 81.72\%, M = 80.1\%$), $t(199) = 1.25, p = .21$, and in end of semester grades, $\chi^2(3, N = 464) = 2.90, p = .406$. A Pearson correlation was conducted to examine the relationship between total quiz points ($M = 597.79$, $SD = 178.10$) and total exam points ($M = 876.97$, $SD = 216.81$). Quiz points were correlated with higher exam points, $r(398) = .71, p < .001$.

There were no significant differences on final course grades ($M = 79.25\%$, $M = 78.82\%$) between the counterbalancing groups, $t(199) = .87, p = .376$.

Discussion

Students performed significantly better on the probe exam questions that related to the quiz sets than they did on questions that were not targeted in the quiz sets. This effect was highly consistent, occurring on each of the 11 unit exams (see Figure 1). The current work demonstrates quizzes need not be administered in class to elicit benefit (cf. Daniel & Broida, 2004), converging with the findings of McDaniel et al. (2012).

Demonstration that quizzing provides a benefit to knowledge within novel contexts (i.e., exam items that are not identical to quiz items) as opposed to a benefit that is limited to multiple presentations of identical items is necessary if the case is to be made that testing promotes knowledge acquisition. Laboratory research has established that transfer of knowledge may occur within testing effect paradigms (Butler, 2010; Carpenter, 2012; Rohrer, Taylor, & Sholar, 2010), but classroom-based research often utilizes identical quiz and exam items (e.g., Carpenter et al., 2009; McDaniel et al., 2011; Roediger et al., 2011), making transfer of knowledge difficult to examine.

McDaniel and colleagues (2013) addressed this issue, demonstrating a testing benefit for both identical and for related nonidentical content. However, this occurred in the context of a middle school classroom rather than at the college level. Additional laboratory research has demonstrated a testing effect for information taken from a college level biology textbook, but only for items that were identical between quiz and exam administration, not for topically related items (Wooldridge et al., 2014). In the current work, we altered question stems and even alternatives when moving from quizzes to exams (see, e.g., Table 2) to determine whether or not students were learning concepts or simply recognizing key words. The presence of a testing effect under these conditions corroborates earlier findings from a middle-school study (McDaniel et al., 2013) and extends them to an ecologically valid high-enrollment college-level course, supporting the idea that quizzes enhance conceptual learning rather than encouraging rote memorization of key words.

While ecological studies are valuable, it is important to minimize penalties to student performance (in the control conditions) when administering classroom experiments for both practical and ethical reasons. Consistent with the Belmont Report (National Commission for the Protection of Human Subjects of Biomedical & Behavioral Research, 1978), which requires researchers to min-

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{Figure1.png}
\caption{Experiment 2: Mean number of questions answered (out of 10 possible) on unit exams as a function of whether the questions corresponded to quizzed items. SEs shown. * $p < .001$.}
\end{figure}
imize harm and maximize benefit to participants, it is desirable that a pedagogical manipulation avoids penalizing students on exams by withholding what we believe to be an effective intervention. We counted correct responses to unit exam questions unavailable to students during quizzing as extra credit. As a result, it was possible to receive a maximum score of 50 out of 40 points (125%) on a unit exam if a student were to correctly answer all of their within-group critical questions (10 points), all of the common questions (30 points), and all of the opposite-group critical questions (10 extra credit points). Despite influencing the total-point score for some students, this manipulation did not affect overall numbers of students receiving a C or better. We believe this is a useful way of comparing pedagogical approaches because it does not carry the potential of penalizing students for performance on “control” items, while still allowing meaningful differences to become apparent.

In Experiments 1 and 2, the benefit of quizzing was confounded with additional exposure to material. There are a small number of studies that have investigated the testing effect within a classroom environment with assessments that count toward student grades; of these, only a subset have controlled for additional exposure by providing a restudy condition (McDaniel et al., 2011, 2013; McDermott et al., 2014; Roediger et al., 2011). To contribute to this literature and determine if exposure as opposed to testing per se provides a plausible explanation for the quizzing benefits found in Experiment 2, we designed an additional experiment to compare quizzing to restudy of information. We hypothesized that quizzing is superior to reading information, even if reading material is presented thematically. To allow the statements to be delivered as quiz items, alternatives were presented to the student as if to gauge their familiarity with the material (their answers were immaterial to the analysis; see Table 2).

Students in Section 002 did not receive statements; they (n = 388) received both their own counterbalancing group and the opposite group’s quiz questions, including critical questions. Each group received the opposite group’s critical questions as statements in Section 001 and as additional quiz questions in Section 002. This design allowed for evaluating replicability of the findings from Experiment 2 and subsequent extension of those findings by inclusion of a course section that used statements embedded in quizzes, enabling both within and between-subjects comparisons of quizzed versus studied information.

Procedure. This research was conducted in compliance with the university’s Internal Review Board. We randomly assigned students to their condition when the instructor received the class list at the start of the semester. Quizzes and unit exams were administered in the same way as described in Experiment 2. At the end of the semester, we give students the opportunity to complete an online, multiple-choice, extra credit survey expressing their attitudes about the course, including what was the best use of their time and which components they liked best.

Results

Quiz performance. Pearson correlations were conducted to examine the relationship between total quiz points and total exam points for each of the two sections. Higher best quiz points (M = 621.82, M = 647.98) were correlated with higher exam points (M = 966.73, M = 994.44) for both sections, (r = .79; r = .80 for Sections 001 and 002, respectively, ps < .001). We conducted t tests to assess if there were any differences between students in Group 1 and students in Group 2 for both sections on the number of times students took the quizzes (M = 6.74, M = 6.42) and best scores on quizzes (M = 15.63, M = 15.81; all ts < 1.0 and all ps > 0.50).

Unit exam performance. A repeated-measures analysis of variance showed that in Section 001, students did better on their quizzed probe questions (M = 6.66, SE = .033) than on probe questions related to their brief statements (M = 5.82, SE = .030), F(1, 403) = 641.50, p < .001, η² = .61. These mean differences were consistent across all exams; all ps < .001 (see Figure 2). There was also no significant difference between students across counterbalancing groups (M = 82.96%, M = 83.24%) on total exam scores, t(201) = 0.823, p = .412, demonstrating that between group performance was similar on all exams.

For Section 002, all students had access to their assigned group’s critical quiz questions and the opposite group’s critical quiz questions. As hypothesized, for Section 002 there were no
differences in performance between same counterbalancing group ($M = 6.49, SE = .033$) and opposite group ($M = 6.48, SE = .032$) probe questions.

Across sections, there were significant differences, however Students in Section 002 did better ($M = 6.40, SE = .031$) on probe questions related to the additional quiz questions they received than students in Section 001 ($M = 5.82, SE = .031$) on the questions related to their brief statements, $F(1, 790) = 172.24, p < .001, \eta^2_g = 0.18$. This effect replicates the within-section results demonstrating the advantage of quizzing relative to restudy of the content from brief statements. Moreover, the between-section analysis allows us to control for item difficulty as we drew Section 001’s restudy statements and Section 002’s extra questions from identical question pools. There were no significant differences on final course grades ($M = 80.81\%$, $M = 78.14$) across sections, $t(1790) = 1.62, p = .107$.

**End of semester survey.** When asked to choose which item represented the best use of their time, the vast majority of students selected taking the quizzes (see Figure 3 for details and proportion of responses to alternatives). When asked what part of the course they liked the most, the most common response was studying by taking quizzes (see Figure 4).

**Discussion**

Both within and between sections analyses demonstrated superiority of quizzing over additional exposure from statements (restudy condition, Section 001). This finding supports both laboratory (Carrier & Pashler, 1992; Karpicke & Blunt, 2011; Wheeler et al., 2003) and classroom-based and online research (McDaniel et al., 2011, 2013; McDermott et al., 2014; Roediger et al., 2011) that demonstrate controlling for amount of exposure fails to account for
the testing effect. Additionally, this effect held even though students were required to respond in multiple-choice fashion to the study statements, therefore controlling for the active component of a motor response inherent in quizzing. The Section 001 within section results replicate those from Experiment 2, suggesting that low-stakes quizzing implemented in the fashion described here produces reliable benefit across semesters.

Students found quizzes to be relatively enjoyable (i.e., they “liked” taking quizzes) and valuable. This finding corroborates those of other studies that have reported generally positive attitudes toward high-frequency, low-stakes testing in a classroom (Bangert-Drouns, Kulik, & Kulik, 1991; Leeming, 2002; McDaniel et al., 2011; Roediger & Karpicke, 2006a). Given that college students tend to list repeated reading of material as their most frequent study method, and perceive restudying information to be the most effective study strategy (Karpicke et al., 2009), it is interesting to note the vast majority of students reported quizzing as more valuable than reading the textbook. We did not survey students regarding their attitudes toward high-frequency low-stakes testing prior to the course, making it difficult to determine if course experience engendered change.

**General Discussion**

Although firmly established in laboratory settings, the effective use of quizzing to enhance learning remains controversial in ecologically valid settings (e.g., Bell et al., 2015; Daniel & Broida, 2004). The aim of the current body of work was to examine the possibility of designing a large introductory college-level course that leveraged research regarding test-enhanced learning. We hypothesized that students would learn better when we gave questions and required students to complete them for credit, thereby combining the benefit of quizzing with the benefit of extrinsic motivation to complete the quizzes. Over the course of several semesters, we were able to demonstrate the advantage of including required frequent, low-stakes quizzing as a graded part of a course.

While a number of researchers have investigated the testing effect within actual classroom settings using course assessments that count toward student grades (Bjork et al., 2014; Lyle & Crawford, 2011; Glass et al., 2008; Glass, 2009; McDaniel et al., 2011, 2012, 2013; McDermott et al., 2014; Roediger et al., 2011), few have included a restudy control condition (McDaniel et al., 2011, 2013; McDermott et al., 2014; Roediger et al., 2011), of which only one occurred in an introductory level college course (McDaniel et al., 2012). As this prior work (McDaniel et al., 2012) used small courses (Ns = 16 and 27), the current work represents a critical extension of prior findings in a novel context under controlled conditions.

The size of the course matters because it provides a great deal of statistical power and, insofar as it relates to demographic and motivational diversity within the course. UNM is an ethnically diverse campus (see Method section for each experiment). At the national level, distribution of undergraduate students enrolled in degree-granting postsecondary institutions is approximately 9% Hispanic, 71% White, 1% American Indian/Alaskan Native, 7% Asian, and 11% Black (National Center for Education Statistics, 2013). As Hispanic students and American Indian students attend UNM at a rate of fourfold and sixfold the national average, respectively, a large course at UNM presents the opportunity to examine the efficacy of quizzing in an introductory college course for these ethnic groups.

With regard to motivational diversity, students in upper level courses are more likely to be prospective majors or minors and exhibit mastery goals (Harackiewicz et al., 2002), but students in an introductory-level course may enroll simply to fulfill basic education requirements. The current study incorporated courses with hundreds of students who approached the course from different motivational and demographic backgrounds. For the wide variety of students enrolled in the course, the consistency of quizzing benefit (see Figures 1 and 2) demonstrated continued effectiveness throughout the semester across the wide variety of topics covered in the course, thereby underscoring the robustness of test-enhanced learning.

In the current work, Experiment 1 indicated an indirect effect of making quizzes a graded course component, as students in the
quiz-required section outperformed students in the quiz-optional section even when controlling for number of quiz attempts and best quiz scores. Experiment 2, however, took place within a single course section in which quizzes were a required course component for all students, thereby controlling for course-engagement due to required quizzes. With this control in place, we found that students learned quizzed information better than nonquizzed information.

Experiment 3 extended these results, indicating structuring information as a question leads to superior learning when compared to information structured as a statement. These findings support the notion that course engagement behaviors are not likely responsible for the entirety of the testing effect. Roediger and Karpicke (2006a) review laboratory research in which retrieval practice elicited benefit within a context that likely lacked motivational components comparable to receiving a course grade. McDaniel, Anderson, Derbish, and Morrisette (2007) reported a testing effect in classroom environments in which the experiment materials did not count toward student grades. Further, Haynie (1997) demonstrated that students who expected but were not presented with a quiz failed to learn as much as students who were not expecting a quiz yet received one. Thus, overall student performance is likely the result of the interplay between direct and indirect contributions of frequent quizzing, and these contributions are not mutually exclusive. However, course engagement behavior alone is unable to explain the testing effect.

Results from Experiment 1 indicate indirect effects of quizzing likely play a role in quizzing benefit within a classroom context, while Experiments 2 and 3 demonstrate that formatting information in the form of a question elicits a benefit. This is consistent with research indicating that some aspect of the retrieval process itself is at the core of the testing benefit, referred to as the retrieval hypothesis (Dempster, 1996). This may occur due to a combination of congruency of processes required at encoding and retrieval (i.e., quizzing and criterial testing, referred to as transfer-appropriate processing; Kolers & Roediger, 1984; Thomas & McDaniel, 2007), elaboration of memory traces (Bjork, 1988; McDaniel, Kowitz, & Dunay, 1989), or increased depth of processing (Karpicke & Blunt, 2011).

In addition, Wing, Marsh, and Cabeza (2013) used functional MRI to examine the brain basis of enhanced recall due to initial testing, revealing regions associated with successful retention include the anterior hippocampus, lateral temporal cortices, and medial prefrontal cortex, as well as increased connectivity between the hippocampus and the ventrolateral prefrontal cortex and midline regions. The authors suggest the testing effect, at the neural level, enhances processing in regions that support successful encoding of material (e.g., relational binding, selection and elaboration of semantically related information) as well as processes related to retrieval, such as memory search.

There are, of course, a number of caveats. We informed students that notes and textbooks were not permitted while taking quizzes, yet it is unknown if and how often such rule violations may have occurred, as the quizzes were administered online. Certain precautions were taken to discourage students from looking up answers (e.g., time allowed for each quiz attempt was constrained, question order and answer option order were randomized for each attempt). Nevertheless, students who consulted books or notes during quiz attempts were not retrieving information in the same fashion as students who completed the quizzes from memory. In Experiment 3, we provided students in both groups with the same information on quizzes, with particular information presented to one group as statements and to the other as questions. If a student cheated by using additional material to look up an answer to a quiz question rather than attempting retrieval of information from memory, this effectively turns that quiz question into a statement (i.e., passive reading of information as opposed to retrieval from memory). Results from Experiment 3 indicate that information posed as a quiz question resulted in superior learning when compared to information in the format of a statement, suggesting that students were largely taking the quizzes as intended. Given that even open-book quizzing can be beneficial (Agarwal et al., 2008), the precautions in place, and the clear benefit of quizzed versus restudied information (see Experiment 3; McDaniel et al., 2012), if students cheated by using additional material during quizzing, it is unlikely it was endemic.

Similar research has demonstrated a testing effect when quizzes were taken in class and spaced at least a day apart, with at least one day between the final quiz and the criterial test (McDermott et al., 2014), as well as when students were allowed to take quizzes whenever they desired, up until one hour prior to administration of a criterial test (McDaniel et al., 2012). The current study allowed students to take the quizzes whenever they wanted (following availability coincident with the first lecture pertaining to unit information relevant to a particular quiz set), including immediately prior to the criterial exams. As we do not have data regarding the timing of quiz completion across days, it is unknown how students tended to space their quiz attempts, and whether this played a role in outcome measures (cf. Cull, 2000).

We did not evaluate duration of learning benefit in the current work. Critical questions were planted within chapter exams, but as students were allowed to take quizzes up until the exam it is unknown when students ceased studying for each exam. Similar research, however, has provided evidence that the testing effect can result in retention benefits of up to 9 months (McDaniel et al., 2011). If an effective learning strategy is based on duration of retention, structuring information in the format of a test is certainly effective (McDaniel et al., 2011; Roediger & Karpicke, 2006b), but the goal of studying may not always be long-term retention.

In any case, students may be motivated primarily to achieve a high grade rather than to retain information over the long term. Reading a great deal of information over a short interval (“crumming”) or taking quizzes immediately prior to an exam may both result in a high grade; the advantage of retrieval practice only becomes evident at long delays following initial learning (Hogan & Kintsch, 1971; Karpicke & Roediger, 2007; Roediger & Karpicke, 2006a; Thompson et al., 1978; Wheeler et al., 2003). Essentially, cramming by reading is effective if the goal is to score well on a test, but not if the goal is to retain learned information, whereas learning via quizzing is effective for both grade achievement and for retention. Students motivated only to score well on a test may therefore require incentive to learn via quizzes (i.e., extrinsic motivation; see Experiment 1). Thus, motivations of students and of instructors (who may want students to perform well in the course, as student achievement may be viewed as reflective of the quality of the instructor) must be taken into account when considering translation from theory to practice.

While instructors may find it difficult to manipulate the motivations of their pupils, there are a number of course parameters
that they may adjust to facilitate test-enhanced learning. We found that providing extrinsic motivation to complete quizzes by making them count toward course grades improved exam performance and grades compared to a similar situation that did not require completion of quizzes (see Experiment 1). Retrieval practice provides a great number of benefits, including enhanced long-term retention (Roediger & Karpicke, 2006a), potentiation of knowledge acquisition on subsequent study trials (Izawa, 1970), identification of knowledge gaps (Dunlosky, Hertzog, Kennedy, & Thiede, 2005), and reduction of test anxiety (Agarwal, D’Antonio, Roediger, McDermott, & McDaniel, 2014; McDaniel et al., 2011). If students cannot be enticed to complete quizzes, they will fail to elicit the benefits provided by quizzing. Requiring quizzes as part of a course grade presumably motivates students to study the material in preparation for the quizzes, learn from the quizzes themselves, and perform well on subsequent tests, thereby supporting course involvement beyond quiz completion (Mawhinney, Bostow, Laws, Blumenfeld, & Hopkins, 1971).

One concern with required quizzes is students will protest at the thought of additional testing, leading to poor ratings of courses that use frequent tests or quizzes. Prior research actually indicates college students prefer courses that include more frequent testing (Bangert-Drowns et al., 1991; Leeming, 2002; Roediger & Karpicke, 2006a; see Experiment 3 of current work, as well). This is likely because increased frequency of testing is associated with lower stakes per test, thereby reducing anxiety and providing students with frequent feedback regarding their course performance (McDaniel et al., 2011). Moreover, our students reported that they found taking quizzes to be a relatively enjoyable course component (see Experiment 3).

The benefits of testing relative to restudy are well established, but testing has not often been compared to other commonly used means of enhancing learning. While testing has been found to be superior to elaborative studying with concept mapping in one instance (Karpicke & Blunt, 2011), testing has not been directly compared to a number of other efficacious teaching methods, such as presentation of novel information or self-explanation (Kornell, Ravelo, & Klein, 2012). Still, testing appears to provide a useful adjunct to teaching, though implementation may be a key factor in eliciting benefit (e.g., providing corrective feedback following each quiz attempt, allowing multiple quiz attempts, reducing factors that discourage retrieval, such as open-book quiz completion; Bell et al., 2015; Daniel & Broida, 2004; McDaniel et al., 2012). The current experiments not only support previous research demonstrating the presence of a testing effect in a classroom environment, but also extend those findings to a large introductory college-level course and suggest a method of evaluating pedagogical interventions in an ecologically valid context.

References


Received June 9, 2015
Revision received January 20, 2016
Accepted January 24, 2016

ORDER FORM
Start my 2016 subscription to the Journal of Experimental Psychology: Applied® ISSN: 1076-898X

Check one:

$64.00 APA MEMBER/AFFILIATE
$145.00 INDIVIDUAL NONMEMBER
$606.00 INSTITUTION

Charge my: □ Visa □ MasterCard □ American Express

Send this order form to American Psychological Association Subscriptions
750 First Street, NE
Washington, DC 20002-4242

For international subscription information, e-mail: subscriptions@apa.org

□ Check enclosed (make payable to APA)

Charge my: □ Visa □ MasterCard □ American Express

Cardholder Name ____________________________

Card No. ____________________________ Exp. Date ____________

Signature (Required for Charge)

Billing Address

Street ____________________________

City ____________________________ State _______ Zip ______

Daytime Phone ____________________________

E-mail ____________________________

Mail To

Name ____________________________

Address ____________________________

City ____________________________ State _______ Zip ______

APA Member # ____________________________

TOTAL AMOUNT DUE $ ______

Sales Tax: 5.75% in DC and 6% in MD

XAPA16