Participation in Voluntary Re-quizzing Is Predictive of Increased Performance on Cumulative Assessments in Introductory Biology

Elise M. Walck-Shannon,[†‡] Michael J. Cahill,[†] Mark A. McDaniel,[†§] and Regina F. Frey[†‖*]

[†]Center for Integrative Research on Cognition, Learning, and Education, Departments of [‡]Biology, [§]Psychological and Brain Sciences, and [‖]Chemistry, Washington University in St. Louis, St. Louis, MO 63130

ABSTRACT

Low-stakes testing, or quizzing, is a formative assessment tool often used to structure course work. After students complete a quiz, instructors commonly encourage them to use those quizzes again to retest themselves near exam time (i.e., delayed re-quizzing). In this study, we examine student use of online, ungraded practice quizzes that are reopened near exam time after a first graded attempt 1–3 weeks prior. We find that, when controlling for preparation (performance in a previous science, technology, engineering, and mathematics [STEM] course and incoming biology knowledge), re-quizzing predicts better performance on two cumulative exams in introductory biology: a course posttest and final exam. Additionally, we describe a preliminary finding that, for the final exam, but not the posttest, re-quizzing benefits students with lower performance in a previous STEM course more than their higher-performing peers. But unfortunately, these struggling students are also less likely to participate in re-quizzing. Together, these data suggest that a common practice, reopening quizzes for practice near exam time, can effectively benefit student performance. This study adds to a growing body of literature that suggests quizzing can be used as both an assessment tool *and* a learning tool by showing that the "testing effect" extends to delayed re-quizzing within the classroom.

INTRODUCTION

Actively engaging students during course work is gaining traction across disciplines and educational levels. However, promoting engagement using carefully crafted activities during class time is just one factor that contributes toward student learning. Increased course structure—for example, providing students with consistent, frequent formative assessments—is another important evidence-based contributor toward effective learning (Freeman *et al.*, 2011; Haak *et al.*, 2011; Eddy and Hogan, 2014). One type of formative assessment often used in highly structured courses is quizzing, or low-stakes testing. Mounting evidence from the laboratory (Roediger and Karpicke, 2006; Karpicke and Roediger, 2008; Karpicke and Blunt, 2011) and the classroom (McDaniel *et al.*, 2012; Orr and Foster, 2013; Bjork *et al.*, 2014; Trumbo *et al.*, 2016) has shown that quizzing does not just serve as an evaluation tool—learning actually occurs *during* quizzing. This phenomenon is referred to as the "testing effect." What is not known, however, is whether there is added benefit to student learning if students retake the quizzes for no additional points before exam time as a study tool. In this study, we ask whether delayed re-quizzing before each exam (used as a study tool and for no additional points) throughout the semester is associated with higher performance on two types of cumulative course assessments in introductory biology and, if

Cynthia Brame, *Monitoring Editor*

Submitted Aug 21, 2018; Revised Jan 17, 2019; Accepted Jan 23, 2019

CBE Life Sci Educ June 1, 2019 18:ar15

DOI:10.1187/cbe.18-08-0163

*Address correspondence to: Regina F. Frey (gfrey@wustl.edu).

so, whether this could simply be explained by greater self-reported study times among students who retook quizzes.

The testing effect has been applied to classroom contexts across the disciplines of psychology (Leeming, 2002; Trumbo *et al.*, 2016), statistics (Lyle and Crawford, 2011), medicine (Larsen *et al.*, 2009; Messineo *et al.*, 2015), and biology (Orr and Foster, 2013; Carnegie, 2015). In an introductory biology classroom, Orr and Foster (2013) studied the use of low-stakes (10% of course grade) quizzes and found that students who participated in 100% of the available quizzes scored better on course exams than students who participated in 0% of the quizzes. In a large, introductory psychology course, Trumbo *et al.* (2016) asked whether quizzing was more effective if it was optional versus a course component. They found that, in both cases, the number of quiz attempts was positively correlated with course grades; however, these quiz attempts were all during the initial learning and not given as a delayed study tool. Classroom studies have also introduced new opportunities to answer social and motivational questions about the testing effect's impact on performance. For example, as Brame and Biel (2015) point out, it is still an open question whether the testing effect is particularly beneficial for certain students, such as those who are underprepared. We hope to add to this growing literature base by asking whether delayed re-quizzing improved student performance and whether these benefits were equal across preparation levels.

To our knowledge, there have been no studies that have implemented a delayed time course of no-stakes re-quizzing for use as a study tool. In this study, we aimed to address whether delayed re-quizzing sufficiently enhanced delayed recall of course content as measured by two cumulative assessments. Our study occurred in a large introductory biology classroom and differed from previous studies in several ways. First, we specified that our re-quizzing attempts would be delayed relative to both the initial learning and initial quiz attempt. Second, the majority of the quiz questions used in this study were randomly chosen from larger pools of highly related questions (see *Methods*); hence, students could see slightly different versions of questions in each of their attempts, and none of the questions used on the quizzes were identical to any questions on either of our outcome variables (i.e., our posttest or final course exam). Third, we controlled for student preparation. Fourth, rather than binning quiz participation into all or nothing, we treated the delayed re-quiz participation as a continuous variable. We hoped that these additional considerations would provide us with a nuanced answer to whether delayed re-quizzing, used as a study tool, was effective in a biology classroom context. This study occupies a specific niche within both the biology education and classroom quizzing literature by analyzing the correlation between delayed re-quizzing and performance on exams that contain completely novel items, while also accounting for students' previous preparatory experiences.

Theoretical Frameworks

The recent classroom applications of quizzing cited earlier were informed by decades of laboratory studies of the testing effect, which have identified factors that enhance or mitigate its magnitude. Here, we will only mention those findings that are relevant to this study; for a more comprehensive review written for biology education researchers, see Brame and Biel (2015). Providing students with feedback (i.e., whether their answer is correct) enhances the testing effect (Butler and Roediger, 2008). In addition, the number of times that a learner practices retrieval for the same prompt, the greater the testing effect, even after successful recall attempts (Roediger and Karpicke, 2006; Karpicke and Roediger, 2008). These laboratory studies suggest that if our goal is to improve semester-long learning, we should provide our students with low-stakes testing opportunities that are sufficiently difficult, occur throughout the semester, and deliver feedback. However, as stated earlier, it is not yet entirely clear whether delayed no-stakes re-quizzing, used as a study tool, has added benefits within a classroom context.

A number of theoretical explanations of the testing effect have been offered in the literature (McDaniel and Masson, 1985; Carpenter, 2009; Pyc and Rawson, 2010; Thomas and McDaniel, 2013). Here, we present two theories that are most relevant to the current study, wherein we examine the benefits of delayed re-quizzing: the current theory of disuse (described by Bjork and Bjork, 1992, as a modification of Thorndike, 1914), and the dual-memory model (Rickard and Pan, 2018).

The current theory of disuse posits that each memory has both a storage strength (i.e., how well a student has learned an item) and a retrieval strength (i.e., how accessible that item is in the current context; Bjork and Bjork, 1992). A student's ability to retrieve information on an exam would be largely related to retrieval strength. While storage strength is theoretically infinite and can increase as we learn more over time, retrieval strength decreases over time as we forget items and are confronted with extraneous information. However, practicing retrieval (in this study, quizzing) has the potential to increase retrieval strength and is especially potent when it is difficult (Bjork, 1994). By having time lapse between students' initial learning of the information and retrieval practice (i.e., quiz first attempt) and later retrieval practice (i.e., quiz retakes), we can essentially make the situation "desirably difficult" for our students. For this reason, we reopened quizzes for voluntary use after a period of 1–3 weeks from initial learning and quizzing, during which the retrieval strength of the content in memory for many students will have presumably decreased.

The second theoretical framework that we considered is the dual-memory model, which posits that testing and restudying can actually form separate memories (Rickard and Pan, 2018). According to the dual-memory model, initial learning or study of a topic yields "study memories" that can be further strengthened by restudying. Retrieval practice, on the other hand, encodes new "test memories" and strengthens existing "study memories." On the final test, students can draw from "study memory," "test memory," or both. Therefore, we would expect a student who has practiced retrieval to have more diverse memories to draw from and to perform better. Notably, in the laboratory, this dual-memory model provides quantitative predictions of the testing effect that could explain the differing results of iterative "restudying" (i.e., rereading) and iterative testing (i.e., re-quizzing consisting of fill-in-the-blank questions) on final test performance (Storm *et al.*, 2014; Rickard and Pan, 2018). Thus, we suspect that re-quizzing using sometimes identical but mostly slightly different items, as we have in the current study, would strengthen "test memory" and encode new "test memories," which would lead to learning gains for our students.

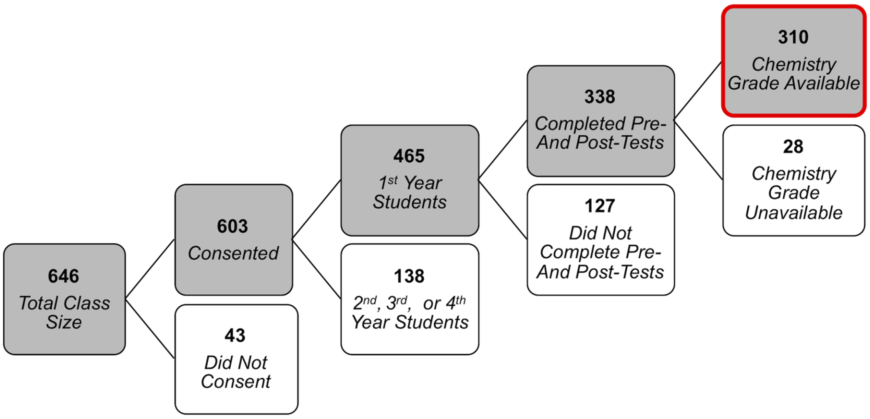
646
Total Class Size
603
Consented
43
Did Not Consent
465
1st Year Students
138
2nd, 3rd, or 4th Year Students
338
Completed Pre- And Post-Tests
127
Did Not Complete Pre- And Post-Tests
310
Chemistry Grade Available
28
Chemistry Grade Unavailable

FIGURE 1. Flowchart of selection criteria for delayed re-quizzing analysis. Each step of exclusion is shown horizontally, and numbers of students at each step are shown in bold. The students represented by the gray (top) box were included for the next step, while the students represented by the white (bottom) box were excluded from the next step. The hierarchical regression models reported include the students represented by the box with a thick red line (*n* = 310). The sample for the study time estimates was further limited by whether or not a student voluntarily self-reported times, as described in the text.

Experimental Questions

Given the overwhelming evidence for the testing effect, both in the classroom and in the laboratory, we wanted all of our students to have some benefit of quizzing. Therefore, we required all students to take one attempt on a weekly quiz, which contributed toward a small part of the course grade. Close to each exam time, the relevant quizzes were reopened for unlimited, ungraded, completely optional practice. This was essentially a “no-stakes” reassessment after a “low-stakes” assessment. Then, at the end of the semester, we measured performance using two different cumulative assessments. Using this design, we aimed to answer the following research question:

> Within a large-enrollment, introductory biology classroom, do students who participate in delayed re-quizzing tend to perform better on cumulative assessments than those who do not participate in re-quizzing when controlling for preparation? And, do students with differing levels of preparation benefit equally from re-quizzing?

On the basis of the aforementioned lab and classroom findings, we hypothesized that the number of quizzes retaken in a delayed manner may predict higher performance on cumulative assessments (i.e., a course posttest and the final course exam). However, we did not have an a priori hypothesis about whether preparation would interact with quizzing. Prior studies have found an equal testing effect over a range of academic abilities (Orr and Foster, 2013), as measured by performance on an earlier course exam.

METHODS

Context and Participants

This study was performed in a large-enrollment (*n* = 646) Introductory Biology I course at a selective, private institution in the Midwest and has been approved by our internal review board (IRB ID#: 201807055). This course covers basic biochemistry, cell signaling, replication, gene expression, and molecular techniques, and it is the first semester of a two-semester sequence. It is typically taken by students interested in life science majors and/or with pre-medicine intentions. We limited our analysis to students who gave consent (93.3%, *n* = 603/646) and first-year students who are in their second semester of college (77.1%, *n* = 465/603). We further limited the sample based on availability of control and outcome variable data, as described below (see *Outcome Variables* and *Selection of Control Variables*) (final *n* = 310, Figure 1). According to registrar records, within this final sample, 39.0% were white (*n* = 121); 41.0% were Asian (*n* = 127); and 17.4% (*n* = 54) were either African American, Hispanic, Native American, or Pacific Islanders, which we grouped together into underrepresented minorities. Ethnicity data were unavailable for eight students. A slight majority of the final sample were women (59.4%, *n* = 184/310).

Quiz Development

To provide students with the opportunity for low-stakes quizzing in this course, an internally funded educational specialist (E.M.W.-S.), who supports faculty as they incorporate evidence-based teaching techniques in the classroom, generated quiz questions, which were iteratively revised in close collaboration with the instructor team for the Introductory Biology I course. We wrote our own quiz questions, rather than using a publisher bank, because when such question banks were used previously in this course, our students reported dissatisfaction due to misalignment with course expectations. To ensure that these questions were better aligned with the course content and level of exam questions, we generated learning objectives for the course, based on exams from previous years, then wrote quiz questions that were aligned with those objectives. Each quiz corresponded to a weekly study guide and ungraded problem set; thus, the quizzes were noncumulative.

The quiz questions were closed response (multiple choice, multiple answer, matching, etc.) so that they could be autograded by our learning management system, which was imperative, given our large class size. About 20% of the exam questions are a similar closed-response format. The quizzes were short, on average 6.23 questions (minimum = 5, maximum = 7, SD = 0.725), and the majority of questions (90%, 72/80) were part of a pool, from which questions were randomly selected for each student. (The remaining 10% of questions were fundamental vocabulary and conceptual comparisons, which were the same for all students and were typically fill-in-the-blank format.) The average number of quiz questions per pool was 4.94 (minimum = 2, maximum = 10, SD = 2.109). The majority of pool questions were different forms of a unified question stem. For example, a student might be asked to determine the coding strand of DNA during transcription. And in different versions, the DNA strands are depicted in different orientations or a different strand is being transcribed. To demonstrate to students that we valued the quizzes, we included one question identical to a quiz question on each of the midterm exams. However, for the tests used as our outcome variables—the final course exam and the posttest—none of the questions was identical to a quiz question.

Quiz Implementation and Analysis

All quiz attempts were given online through our learning management system. Each week, students were assigned a quiz to take for credit (i.e., first attempt). The highest 11 (out of 13) quizzes counted toward 10% of the total grade. The vast majority of the students in our final sample completed all of the quizzes required for the course (98.7%, n = 306/310). In the week before exam time, quizzes related to that exam were reopened to students for unlimited, untimed, ungraded practice (i.e., delayed no-stakes re-quizzing). Students received an email announcement when the quizzes had been reopened for the purpose of studying for the exam. Before exam 1, this opportunity was also announced in class. Feedback during delayed re-quizzing was provided immediately after completing each attempt, whereas feedback after the graded first attempt was released after the quiz due date to minimize answer sharing. Feedback included the student's answer, the correct answer, and in some cases, an explanation for why a particular answer was correct for a specific question. Because each midterm exam was noncumulative, quizzes were only reopened once, right before exam time. A total of 10 quizzes were reopened for practice during the semester: four before exam 1 (9 weeks before the cumulative assessments), three before exam 2 (5 weeks before the cumulative assessments), and three before exam 3 (10 days before the cumulative assessments).

For the purposes of this study, we analyzed only the ungraded practice attempts (i.e., the delayed re-quizzing) and not first attempts, which were required for the course grade. To determine whether a student retook a quiz for the purpose of studying for an exam, we retrieved the quiz-attempt log from our learning management system, then counted the number of available delayed quizzes each student retook at least once, which ranged from 0 to 10. For simplicity, in this study, we considered only whether a student retook each delayed quiz and not the number of attempts per quiz or the score of each attempt in the delayed opportunity.

Outcome Variables

To assess the effectiveness of re-quizzing, we analyzed two cumulative assessments: the course posttest and the final course exam.

Posttest. Our primary outcome measure was the course posttest (Supplemental Material 1), which was a multiple-choice measure developed by our instructor team to align with the course goals and measure learning gains in Introductory Biology I. Students take this test within the first week of the semester ("pre," see *Selection of Control Variables*) and the last week of the semester ("post"), which allows for within-subject knowledge comparisons. The test was first developed 6 years before the present study (2013) and has been iteratively improved based on difficulty and discrimination measures from student responses. The mean of the posttest was 26.74 (70.37%) with an SD of 6.26 (full descriptive statistics are reported in Supplemental Material 2). To verify that the test aligned with the course focus, we qualitatively categorized questions by topic and compared their distribution to the course syllabus. Two raters categorized questions (interrater kappa was 0.94), then met to discuss discrepancies until agreement was reached. The distribution of these agreed-upon topic categorizations did not significantly differ from the distribution of topics covered during lecture as determined by the course syllabus ($\chi^2 = 58.95$, $p = 0.96$), suggesting that the content coverage was appropriate for this course. Because the internal reliability was within acceptable range (Cronbach's alpha for all items = 0.73), the scores on all items were averaged together into a single score for each student, which is reported here.

Similar to the quizzes, the pre- and postcourse tests were completed outside class on our learning management system. Students received a small amount of credit for completing the pre and posttests (low-stakes). While the accuracy of their answers did not count toward the course grade, students were told that "reasonable effort" would be required for credit and were encouraged to view the posttest as an opportunity to review course material and identify concepts for further study.

Final Course Exam. We also conducted secondary analyses, in which the final course exam was used as the outcome variable. The majority of this exam is cumulative and was written collaboratively by the instructor team. The exam was given in class and accounted for 20% of the course grade (i.e., high-stakes). The mean of the exam was 84.33 with an SD of 9.93 (full descriptive statistics are reported in Supplemental Material 2). The internal reliability was within acceptable range (Cronbach's alpha = 0.84). Because this is not a validated research instrument and changes from year to year, we did not consider this to be a primary outcome. However, after examining the effects of delayed re-quizzing on the posttest, we conducted secondary analyses to see whether the general pattern of effects also held for the final course exam. One consideration in conducting these analyses is that effects on the cumulative final exam, which counts toward course grades, may resonate more with students and instructors than effects on the ungraded posttest alone.

Selection of Control Variables

Potential Preparatory Control Variables. Given our experimental questions, we aimed to control for student preparation. To select which preparation variables to include, we looked at the pairwise correlations between the primary outcome performance variable (i.e., the posttest score) and the following available preparatory scores.

College Preparation. We examined two measures of college readiness: ACT scores and Advanced Placement (AP) success in science, technology, engineering, and mathematics (STEM). We found a small but significant correlation between ACT composite score and posttest score ($r(281) = 0.23$, $p < 0.0001$). For our AP success measure, we calculated the proportion of AP course work from four STEM disciplines (biology, chemistry, calculus, and physics) that a student successfully completed. A score of 4 or 5 on the AP exam was considered to be successful completion. Using this AP proportion measure, we found a significant moderate correlation with biology posttest score ($r(308) = 0.37$, $p < 0.0001$; Cohen, 1992).

Incoming Biology Knowledge. To gauge students' incoming biology content knowledge, we used the course pretest, which contains the same items as the posttest but is given during the first week of class. Not surprisingly, we found a significant moderate correlation between the pretest and posttest ($r(308) = 0.43$, $p < 0.0001$).

Previous STEM Course Performance. Because first-year students take Introductory Biology I in their second semester, we also have scores on their previous STEM course work that could be used as preparation variables. The vast majority of students take the first semester of general chemistry in the semester preceding Introductory Biology I. Owing to the limited overlap in content of General Chemistry I and Introductory Biology I, we suspect that a correlation would result from general cognitive and noncognitive skills (critical thinking, problem-solving skills, social skills, persistence, etc.) that students are able to apply across college-level STEM courses. (On the other hand, the biology pretest would include content-specific knowledge and cognitive skills.) When using cumulative final exam scores from General Chemistry I, we found a significant large correlation to biology posttest score ($r(308) = 0.50$, $p < 0.0001$; Cohen, 1992).

Selected Preparatory Control Variables. As stated earlier (and summarized in Supplemental Material 3A), we found the highest correlation with posttest score and the two preparatory control variables: a previous STEM course final exam score and the biology pretest score. While it is straightforward that the biology pretest represents incoming biology knowledge, we suspected that previous STEM course work and college preparation measures may have represented somewhat overlapping attributes.

To check this interpretation, we asked whether the proportion of successful STEM AP course work and ACT composite score added any unique explanatory power to a model predicting posttest score. We used the following hierarchical approach for this analysis. Once the biology pretest and first-semester General Chemistry I final exam scores were incorporated into a model predicting posttest scores, adding in both the AP STEM proportion and the ACT composite score did not significantly increase the explanatory power of the model ($F(1, 278) = 0.2651$, $p = 0.7673$, $\Delta R^2 = 0.0013$; Supplemental Material 3B). Therefore, we limited our preparatory variables to biology pretest score and final exam score in a previous STEM course (General Chemistry I).

Potential Personality Control Variables. We also had access to a subset of students' personality traits, as measured by the Big Five Inventory (Goldberg, 1993; John *et al.*, 2008) one semester prior, but these were not significantly correlated with posttest or final exam score (Supplemental Material 3A). Additionally, once the biology pretest and first-semester chemistry final exam scores were incorporated into a model predicting posttest scores, adding in conscientiousness did not significantly increase the explanatory power of the model ($F(1, 227) = 0.907$, $p = 0.3418$, $\Delta R^2 = 0.0023$; Supplemental Material 3C). Hence, we did not use these personality trait measures as predictor variables in our models.

Final Sample. In summary, we decided to include the final exam grade from a previous STEM course (General Chemistry I) and the biology pretest scores as our control variables. Therefore, students included in this study were first-year students who had completed every question on the pre- and posttests and for whom we had General Chemistry I and Introductory Biology I final exam grades (Figure 1, $n = 310$), unless otherwise noted.

Statistical Method

For this analysis, we used a form of multiple regression, a method that has been suggested for use by biology education researchers to more accurately estimate the effects in their studies (Theobald and Freeman, 2014). The method we used, called hierarchical multiple regression, sequentially adds groups of predictor variables to the statistical model so that one can observe whether a specific predictor or group of predictor variables significantly improves the model's explanation of the outcome variable when other predictor variables have already been accounted for (such as preparation). Because we know that all readers may not be familiar with this method, we insert some explanatory details in the following sections.

Specifically, we report two main hierarchical models and two post hoc models. The first model predicts the primary outcome variable of posttest score and the second and post hoc models predict the secondary outcome variable of final exam score. These models incorporate preparation, delayed re-quizzing, and preparation by delayed re-quizzing interactions as predictor variables in steps 1, 2, and 3, respectively. Unless otherwise noted, standardized β values are reported for each predictor variable, which represent the number of standard deviations of change in the outcome variable (i.e., posttest or final exam score) that would result for every 1 SD change in the predictor variable of interest. Thus, these standardized β values allow us to weigh the importance of individual predictor variables toward explaining the outcome variable. An *F*-test was performed for each hierarchical step to determine whether the fit of the model was significantly improved by the addition of the predictors included in that current step, relative to the previous step, more than is expected by chance. R was used for the hierarchical regression analyses (R Core Team, 2013) and JMP (v. 14; SAS Institute, Cary, NC) was used for all remaining analyses.

Self-Reported Study Time. Additionally, at six times throughout the semester, we asked students to report the number of hours that they spent studying in the previous week. Three of these reports were on exam weeks and three were on non–exam weeks. Reporting study time was completely voluntary for students and was gathered on quiz first attempts. Students were included in the analysis of study time if they made any number of reports (1–3). If multiple reports existed, they were averaged across exam or non–exam weeks. Numbers above 3 SD of the mean were excluded from the analysis.

Delayed Re-quizzing Attempt Times

To determine the mean duration of re-quizzing attempts, we obtained raw attempt times from the quiz-attempt log. This includes every delayed re-quizzing attempt for every student. To account for re-quizzing attempts during which browsers were left open for long periods, we excluded attempt times greater than double the intended time (i.e., greater than 30 minutes), which represents <5% of all recorded attempts ($n = 222/5516$).

RESULTS

Participation in Delayed Re-quizzing throughout the Semester Predicted Higher Scores on the Low-Stakes Cumulative Posttest

We performed hierarchical multiple regression first to predict our primary outcome measure: posttest scores. Our first hierarchical regression model was aimed at answering our research question, namely, does delayed re-quizzing increase performance on cumulative assessments when controlling for preparation and does re-quizzing help all preparation groups equally? Therefore, we added our predictor variables into the model in three steps: 1) preparation control variables, 2) delayed re-quizzing participation, and 3) interaction terms between delayed re-quizzing and preparation.

Model 1, Step 1: Preparation. The first step of our hierarchical regression analysis predicted posttest score from two preparation variables, pretest score and final exam score in a previous STEM course (General Chemistry I; see *Methods* section for justification). Despite having the same questions on the pretest and posttest, in this case, the final exam score in a previous STEM course (General Chemistry I; $\beta_{std} = 0.404$; step 1, Table 1) was even more informative for predicting the biology posttest score than the score on the biology pretest ($\beta_{std} = 0.309$; step 1, Table 1). Together, these two preparation terms explained 33.5% of the variance in biology posttest score.

Model 1, Step 2: Delayed Re-quizzing. We next asked whether adding in re-quizzing participation significantly improved the explanatory power of the model that already accounted for student preparation. Therefore, in step 2, we added to the model the number of delayed quizzes that a student retook at least once (0–10). For example, a value of 0 denotes that only the first, graded attempts were completed during the semester and a value of 10 denotes that a student retook every delayed quiz available at least once after the first attempt. We found that there was a significant increase in the explanatory power of our model when adding in the delayed re-quizzing variable ($F(1, 306) = 8.790$, $p = 0.003$, $\Delta R^2 = 0.018$; see also step 2 of Table 1). When looking at the individual predictor terms, delayed re-quizzing was significant ($p = 0.0034$) but less predictive than either preparatory variable in predicting posttest score ($\beta_{std\ re\text{-}quiz} = 0.139$ vs. $\beta_{std\ pretest} = 0.329$, $\beta_{std\ chem} = 0.374$; step 2, Table 1). The unstandardized estimate, or the estimate in its original units, of the delayed re-quizzing effect was 0.233. This means that, on average, for every quiz that a student retook, he or she scored 0.6% higher on the posttest (i.e., 0.233 gain out of a total of 38 points). Therefore, on average, a student who retook all 10 quizzes scored 6% higher on the posttest than a student who retook none. Taken together, these results suggest that delayed re-quizzing promoted a small but significant increase in posttest scores among our students.

Model 1, Step 3: Delayed Re-quizzing Interactions with Preparation. Next, we sought to determine whether this delayed re-quizzing effect was equal across preparation levels. Because we did not have an a priori prediction of whether delayed re-quizzing would help students with differing incoming biology knowledge (pretest score) or differing success in previous STEM course work (chemistry final exam score), we added both interaction terms into our model in step 3. This step did not significantly increase the explanatory power of the model ($F(2, 304) = 2.421$, $p = 0.091$, $\Delta R^2 = 0.010$; see also step 3, Table 1). The coefficient for the chemistry final exam by re-quizzing interaction was significant, but because the overall step did not improve the model, we refrain from interpreting this interaction at this time.

In summary, even when controlling for preparation, delayed re-quizzing was an effective learning tool for our students as measured using a low-stakes biology posttest. And, according to this analysis, there was no robust evidence that the influence of delayed re-quizzing was dependent on student preparation. Though the values differed slightly, these same patterns were observed when the model also accounted for scores on the first, graded quiz attempts (Supplemental Material 4A).

Participation in Delayed Re-quizzing throughout the Semester Predicted Higher Scores on a High-Stakes Cumulative Final Course Exam

Given the observed effect of delayed re-quizzing on posttest scores outlined in Table 1, we sought to use a second outcome variable to look for similar patterns. For this second statistical

TABLE 1. Standardized β values for each predictor variable (β_{std}) and model-level statistics of hierarchical regression analysis that predicts biology posttest score[a]

	Standardized predictor coefficients[b]		
	Step 1	Step 2	Step 3
Chemistry final exam score	0.404***	0.374***	0.353***
Biology pretest	0.309***	0.329***	0.329***
Number of quizzes retaken	–	0.139**	0.158***
Chemistry final exam score × number of quizzes retaken	–	–	–0.107*
Biology pretest × number of quizzes retaken	–	–	0.011
	Model-level statistics[c]		
R^2	0.335	0.353	0.363
R^2_{adj}	0.331	0.347	0.353
ΔR^2	–	0.018**	0.010

[a] The following symbols indicate significance: *$p \le 0.05$; **$p \le 0.01$; ***$p \le 0.001$; $n = 310$.

[b] Please see the *Methods* section for a description of standardized β values.

[c] For the model-level statistics, an *F*-test was performed to determine whether each step significantly improved the fit of the model relative to the previous step.

TABLE 2. Standardized β values for each predictor variable (β_{std}) and model-level statistics of hierarchical regression analysis that predicts cumulative final exam score[a]

	Standardized predictor coefficients[b]		
	Step 1	Step 2	Step 3
Chemistry final exam score	0.565***	0.545***	0.516***
Biology pretest	0.265***	0.277***	0.277***
Number of quizzes retaken	–	0.090*	0.115**
Chemistry final exam score × number of quizzes retaken	–	–	–0.138**
Biology pretest × number of quizzes retaken	–	–	–0.022
	Model-level statistics[c]		
R^2	0.480	0.488	0.508
R^2_{adj}	0.477	0.483	0.500
ΔR^2	–	0.008*	0.020**

[a] The following symbols indicate significance: $*p \le 0.05$; $**p \le 0.01$; $***p \le 0.001$; $n = 310$.

[b] Please see the *Methods* section for a description of standardized β values.

[c] For the model-level statistics, an *F*-test was performed to determine whether each step significantly improved the fit of the model relative to the previous step.

model, we used the course final exam as the outcome variable rather than posttest score, but we used the same steps (preparation, delayed re-quizzing participation, and preparation by delayed re-quizzing) as our first model.

Model 2, Step 1: Preparation. First, we used the preparation variables of final exam score in previous STEM course work (General Chemistry I) and biology pretest score to explain biology final exam performance. As in model 1 explaining the biology posttest, in model 2, we found a greater effect of chemistry final exam score on biology final exam performance than biology pretest score ($\beta_{std} = 0.565$ vs. 0.265; step 1, Table 2). This large effect of chemistry final exam score makes sense given the similar assessment types of these two variables (high-stakes traditional testing situations). We found that these two preparatory factors, when taken together, explained 48.0% of the variance in the final biology course exam score.

Model 2, Step 2: Delayed Re-quizzing. Next, we added delayed re-quizzing participation throughout the semester into our model at step 2. This slightly, but significantly, improved the model's explanatory power in predicting the biology final exam ($F(1, 306) = 4.813$, $p = 0.029$, $\Delta R^2 = 0.008$; see also step 2, Table 2). Put in more absolute terms, the unstandardized estimate of delayed re-quizzing was 0.240, meaning that, on average, for every quiz that a student retook, he or she scored 0.24% higher on the final exam (which is out of a total of 100 points). Therefore, we would predict that a student who retook all 10 quizzes would score 2.4% higher on the final exam than a student who retook none. Further, we found that the effect of delayed re-quizzing was much smaller than the preparation variables ($\beta_{std} = 0.090$; step 1, Table 2). Taken together, these results suggest that delayed re-quizzing promoted a small but significant increase in high-stakes final exam scores among our students.

Model 2, Step 3: Delayed Re-quizzing Interactions with Preparation. Finally, we asked whether the effect of delayed re-quizzing on final exam scores was equal across preparation levels by adding the same two preparation by re-quizzing interaction terms into our model at step 3. Adding in these terms significantly improved the explanatory power of our model predicting final exam score ($F(2, 304) = 6.251$, $p = 0.002$, $\Delta R^2 = 0.020$; see also step 3, Table 2). Additionally, we found a negative interaction with re-quizzing and chemistry exam score; that is, delayed re-quizzing was most helpful for students who previously struggled in a STEM course (General Chemistry I). But there was not a significant relationship between incoming biology knowledge (pretest score) and re-quizzing (step 3, Table 2). To put this in more absolute terms, on average, a student who scored in the 25th percentile for both the chemistry final exam (75/120) and biology pretest (14/38) would earn 76.5% on the biology final if they did not participate in re-quizzing but would earn 82.0% on the biology final if they re-quizzed at every opportunity. On the other hand, on average, a student who scored in the 75th percentile for both the chemistry final exam (97.25/120) and biology pretest (22/38) would earn 90.1% on the biology final if they did not participate in re-quizzing and would earn 90.5% on the biology final if they re-quizzed at every opportunity. This can be seen graphically in Figure 2A, as students participate more in re-quizzing (*x*-axis), there is less of an effect across chemistry exam score (i.e., there is a decrease in the slope across the panels in Figure 2A). Further, when we looked back at a similar plot for the posttest (Figure 2B), the pattern still existed, though it was not at the level of statistical significance. Conversely, there was no significant difference in the effectiveness of re-quizzing depending on students' pretest score (step 3, Table 2). Hence, when considering biology final exam scores, delayed re-quizzing is more effective for students who have struggled in previous STEM course work than those who have been more successful.

In summary, when controlling for preparation, delayed re-quizzing was an effective learning tool, as measured by the biology cumulative final exam scores. According to this analysis predicting the high-stakes final, there was a significantly greater benefit of delayed re-quizzing for students with lower final exam scores in a previous STEM course (General Chemistry I). Though the values differed slightly, we also observed a significant interaction with preparation and delayed re-quizzing in a model that also accounted for scores on the first, graded attempts (Supplemental Material 4B).

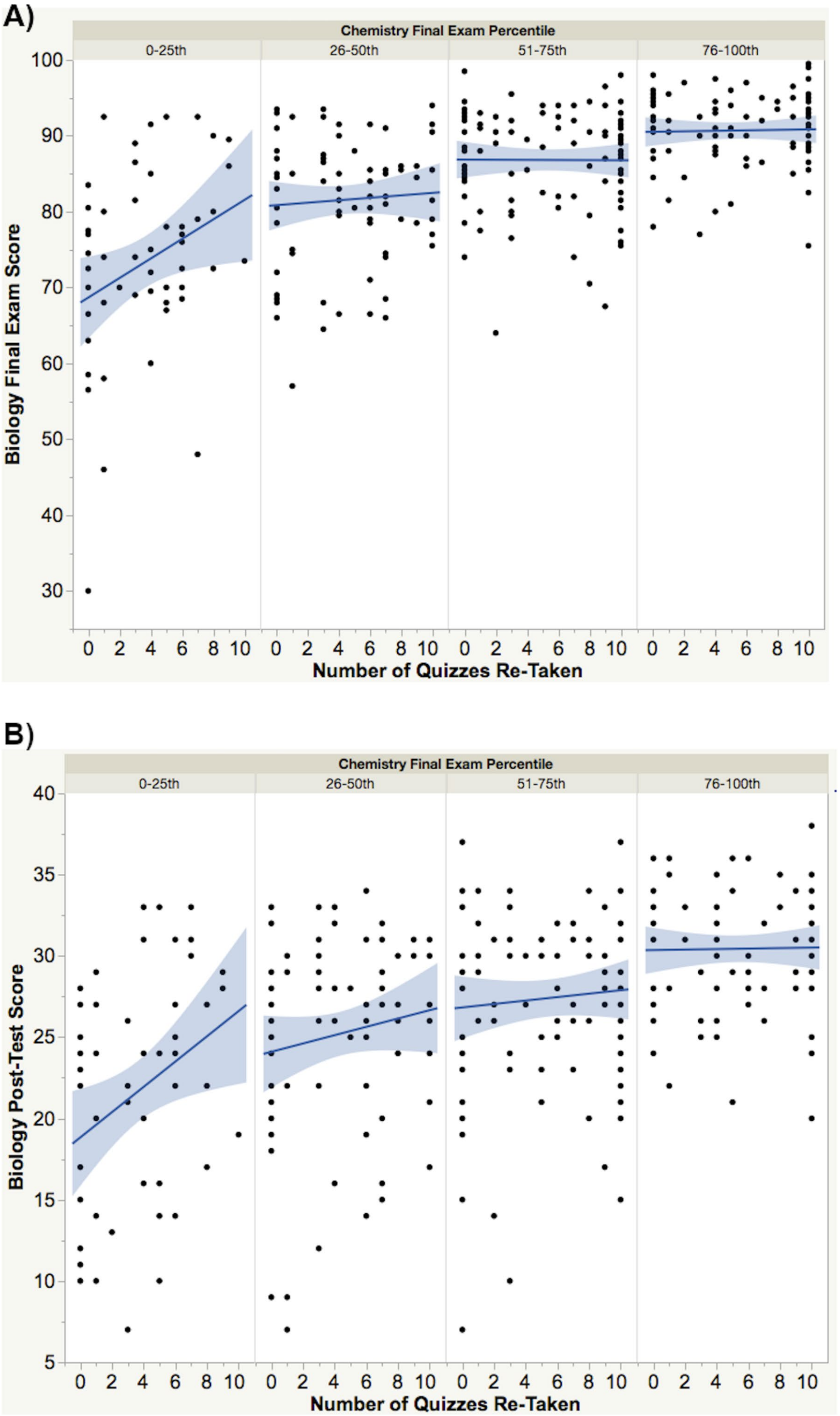

FIGURE 2. Scatter plots relating biology final exam score (A) or biology posttest score (B) to the number of quizzes retaken during the semester. Each graph is split into four panels that represent quartiles for the final exam score in a previous STEM course (General Chemistry I). Blue line indicates best-fit, while the shaded area around it represents the 95% confidence interval for the fit.

Models 3 and 4: Post Hoc Investigation of Exam Question Type. After observing a relationship between delayed re-quizzing and biology final exam scores, we next decided to examine whether the correlation was specific to closed response (multiple choice, matching, etc.) and/or free-response question types as a post hoc analysis. To do so, we calculated closed-response (37/100 points) and free-response (63/100 points) subscores from each student's final exam, then performed the same hierarchical regression analysis described earlier separately with each subscore. We observed complementary results when predicting closed- and free-response final subscores.

When predicting closed-response subscores (model 3, Table 3A), we found that, even after accounting for student preparation (step 1), the addition of delayed re-quizzing (step 2) significantly increased the explanatory power of the model ($F(1, 306) = 6.217$, $p = 0.010$, $\Delta R^2 = 0.012$). However, the addition of preparation interactions into this model did not improve its explanatory power ($F(2, 304) = 2.219$, $p = 0.110$, $\Delta R^2 = 0.009$; Table 3A). Put in more absolute terms, the unstandardized estimate of re-quizzing was 0.119, meaning that we would predict that a student who retook all 10 quizzes would score 3.2% (1.19/37 total points) higher on the closed-response question component of the final exam than a student who retook none.

On the other hand, when predicting free-response subscores (model 4, Table 3B), we found that, after accounting for student preparation (step 1), the addition of delayed re-quizzing (step 2) did not significantly increase the explanatory power of the model ($F(1, 306) = 6.217$, $p = 0.010$, $\Delta R^2 = 0.012$). Interestingly, however, the addition of preparation and delayed re-quizzing interactions (step 3) did significantly increase the explanatory power of this model ($F(2, 304) = 7.276$, $p = 0.001$, $\Delta R^2 = 0.027$). That is, delayed re-quizzing tended to have the greatest benefit in the biology free-response final exam scores among students with the lowest General Chemistry I scores, with the delayed re-quizzing benefit decreasing as students' General Chemistry I scores increased (see graph in Supplemental Material 5). To put this in more absolute terms, on average, a student who scored in the 25th percentile for both the chemistry final exam (75/120) and biology pretest (14/38) would earn 77.4% (48.74/63) on the free-response portion of the biology final if they did not participate in delayed re-quizzing but would earn 83.0% (52.27/63) if they re-quizzed as a study tool at every opportunity; whereas, on average, the model would not

TABLE 3. Models 3 and 4 (post hoc): standardized β values for each predictor variable (β_{std}) and model levels statistics of hierarchical regression analysis that predicts closed-response (A) and free-response (B) portions of the cumulative final exam separately[a]

A. Closed-response portion of final (model 3)			
	Standardized predictor coefficients[b]		
	Step 1	**Step 2**	**Step 3**
Chemistry final exam score	0.486***	0.462***	0.446***
Biology pretest	0.244***	0.260***	0.260***
Number of quizzes retaken	–	0.115**	0.125**
Chemistry final exam score × number of quizzes retaken	–	–	–0.055
Biology pretest × number of quizzes retaken	–	–	–0.062
	Model-level statistics[c]		
R^2	0.369	0.381	0.390
R^2_{adj}	0.365	0.375	0.380
ΔR^2	–	0.012**	0.009

B. Free-response portion of final (model 4)			
	Standardized predictor coefficients[b]		
	Step 1	**Step 2**	**Step 3**
Chemistry final exam score	0.530***	0.517***	0.484***
Biology pretest	0.249***	0.258***	0.257***
Number of quizzes retaken	–	0.060	0.091*
Chemistry final exam score × number of quizzes retaken	–	–	–0.170**
Biology pretest × number of quizzes retaken	–	–	0.011
	Model-level statistics[c]		
R^2	0.424	0.427	0.454
R^2_{adj}	0.420	0.422	0.445
ΔR^2	–	0.003	0.027***

[a] The following symbols indicate significance: *$p \le 0.05$; **$p \le 0.01$; ***$p \le 0.001$; $n = 310$.

[b] Please see the *Methods* section for a description of standardized β values.

[c] For the model-level statistics, an *F*-test was performed to determine whether each step significantly improved the fit of the model relative to the previous step.

predict a meaningful difference for a student who scored in the 75th percentile for both the chemistry final exam (97.25/120) and biology pretest (22/38). Such a student would be predicted to earn 92.4% (58.21/63) on the free-response portion of the biology final if they did not participate in delayed re-quizzing and would earn 91.4% (57.57/63) if they re-quizzed as a study tool at every opportunity.

Together, these results recapitulate the results from the entire final, but suggest that the effects of delayed re-quizzing are unique for different questions types: delayed closed-response re-quizzing significantly predicted higher closed-response final exam scores regardless of success in previous course work; whereas, for the free-response questions on the biology course final, delayed re-quizzing particularly benefited students who struggled in previous course work.

Success in Previous STEM Course Work Is Correlated with Participation in Delayed Re-quizzing

Given this potential pattern that delayed re-quizzing may be more helpful for students who have struggled in the previous chemistry course, we wanted to investigate whether such students were taking advantage of the opportunity. Therefore, we looked at the number of quizzes that students retook in relation to their chemistry final exam grade and found a weak but positive correlation (pairwise $r(310) = 0.170$, $p = 0.002$). In other words, students who were more successful in a previous STEM course were more likely to take advantage of the delayed re-quizzing. This suggests that, despite the potential trend that delayed re-quizzing is especially helpful to students who have scored lower in the previous chemistry course, these students are significantly less likely to participate.

Students Who Participated in Delayed Re-quizzing Did Not Report More Time Studying

Finally, we questioned whether students who participated in delayed re-quizzing were just studying more. To answer this question, we relied on self-reported study times gathered at multiple points throughout the semester. We separated these self-reports into exam weeks, when re-quizzing was available, and non–exam weeks, when re-quizzing was not available. We wanted to check the latter, because if re-quizzing participation correlated with study time even when it was not available, this would suggest that there was some confounding difference between re-quizzing participants and nonparticipants. In exam weeks, students reported studying 4.5 hours on average; while in non–exam weeks, students reported studying nearly 4 hours on average (Table 4). We observed that delayed re-quizzing participation was not correlated with study times in exam weeks (pairwise $r(237) = 0.0274$, $p = 0.675$) or non–exam weeks (pairwise $r(229) = –0.035$, $p = 0.602$). These results suggest that the extent of delayed re-quizzing participation does not affect students' own estimates of their study time.

TABLE 4. Self-reported study time and re-quiz durations with mean durations and SDs shown in hours, minutes, and seconds (hh:mm:ss)

	Mean (hh:mm:ss)	SD (hh:mm:ss)
Time studying per non–exam week	04:08:04	02:38:08
Time studying per exam week	04:30:18	03:08:01
Time per re-quizzing attempt	00:05:32	00:04:16
Total time of all re-quizzing attempts per student	01:02:44	00:52:51

We further asked whether this interpretation was reasonable by objectively observing the time commitment for re-quizzing based on attempt logs. The average re-quizzing attempt duration was close to 5 minutes (Table 4). Because students may have attempted a delayed re-quiz more than once, we also summed all re-quiz attempt durations from all available quizzes across the semester for each student. On average, for students who re-quizzed, this summed duration over the course of the semester was about 1 hour (Table 4). This suggests that delayed re-quizzing requires only a small time commitment from the student and does not constitute a significant portion of students' weekly study-time estimates.

DISCUSSION

Was Delayed Re-quizzing Effective?

In this work, we found that no-stakes, delayed re-quizzing was an effective learning tool in a large classroom context as measured by two cumulative assessments, biology knowledge posttest and biology-course final exam score. This finding is particularly meaningful, given that neither of these cumulative assessments contained questions identical to quiz questions, but rather asked different questions based on the same concepts. The posttest was low-stakes—it was graded only on completion and was implemented outside class—meanwhile, the course final was high-stakes—it constituted 20% of the course grade and was given in class. These assessments are also different in structure, with the posttest being completely closed-response multiple-choice questions and the final being a mix of closed-response and free-response questions. In fact, when we separated out the closed- and free-response portions of the final exam in a post hoc analysis, we found that closed-response delayed re-quizzing was helpful for all students—regardless of preparation—on the closed-response portion of the final exam, whereas re-quizzing was particularly helpful for students with lower success in a previous STEM course on the free-response portion of the final exam. This finding differs from another classroom study (McDermott *et al.*, 2014) in which the testing effect was largely independent of question type, but is consistent with the majority of laboratory findings that the magnitude of the testing effect depends on question type (Kang *et al.*, 2007; Rowland, 2014; see Smith and Karpicke, 2014, for an exception). Together, our observations using these fundamentally different assessments suggests that the positive effect of delayed re-quizzing holds over different performance pressures; however, in our analysis, the main effect of delayed re-quizzing was specific for questions of a similar type on the course final and posttest.

The result that delayed re-quizzing was correlated with higher cumulative exam scores is consistent with theoretical views of the testing effect presented in the *Introduction*. On the basis of the current theory of disuse, we designed a situation where there was a delay between initial learning/quizzing and the retake attempts. On the basis of the dual-memory theory, we designed each quiz attempt to differ slightly for the majority of questions so that "test memories" would be both formed (novel questions) and strengthened (identical questions to previous attempts) and "study memories" would be strengthened. One interpretational limit to our study, however, is that the delayed timing of the re-quizzing (i.e., the "placement" relative to the cumulative assessment) could be the major mediator of the observed effect, rather than the fact that these were re-quizzing attempts. However, it is important to note that, in the present study, the end-of-semester cumulative assessments were administered as long as 9 weeks after re-quizzing. Hence, the present findings are perhaps more likely reflective of benefits of re-quizzing rather than quiz placement per se. Still, future studies are needed to further disentangle these two factors.

Strength of the Testing Effect for Delayed Re-quizzing in Our Study

It is worth noting that the delayed re-quizzing effect sizes when predicting either the posttest ($\beta_{std} = 0.1581$) or the course final exam ($\beta_{std} = 0.1149$) were small in size (but still within the expected range) compared with previous studies examining the testing effect (compiled in Rickard and Pan, 2018). We suggest a few potential explanations for this. First, some of the items on the posttest and cumulative final covered concepts that were not covered on the reopened delayed quizzes. Second, unlike many previous studies examining the testing effect, we did not include identical questions in the quizzes and cumulative assessments (i.e., our outcome variables). In previous studies, even small differences between quiz and exam questions decreased (McDaniel *et al.*, 2012) or completely mitigated (Wooldridge *et al.*, 2014) the testing effect. Third, and perhaps most importantly, students may have tested themselves in contexts outside the delayed re-quizzing opportunities that we created. For example, students who did not retake the quizzes online may still have tested themselves using flashcards or our course-organized peer-learning teams. If these active approaches were used by students who did not retake quizzes, they could have possibly lowered the magnitude of the testing effect. Even though the effect sizes of delayed re-quizzing that we observed were small, they were significant and required a relatively small time investment on the part of both the student (Table 4) and the instructor; furthermore, they may have particularly helped vulnerable students who had struggled with previous STEM course work.

Explanations for the Effect of Delayed Re-quizzing beyond the Testing Effect

Given the classroom context of this study, we—like others (McDaniel *et al.*, 2013; McDaniel and Little, 2019)—appreciate that factors other than the act of retrieval practice itself likely contribute to the correlation that we observed between delayed re-quizzing and exam grades. For example, re-quizzing may help students metacognitively monitor their knowledge so that they can follow up with further studying. While we did not observe any increase in self-reported study time depending on the number of quizzes retaken, it may be that students who re-quizzed were able to better focus their study time based on the feedback of those attempts.

Did the Benefits of Delayed Re-quizzing Depend on Preparation?

We observed mixed results while trying to investigate whether there are differing effects of delayed re-quizzing for different preparatory levels. While predicting both the posttest (model 1) and course final exam (models 2–4), we noticed an interesting trend, wherein students who struggled in a previous STEM course (General Chemistry I) benefited more from delayed re-quizzing than their peers who had higher previous achievement in this STEM course.

The analysis using the posttest hinted at this interaction, but the statistical support was equivocal, given that step 3 of model 1, which included this interaction term, did not significantly improve the model. On the other hand, the analysis using the entire final exam (model 2) or the free-response portion of the final exam (model 4) shows strong statistical support for this interaction, with the drawback being that the final exam score was a secondary outcome measure, because it was not intended for research. However, strong support for this interaction comes from the fact that both posttest and final exam analyses converge, in an almost identical manner, on a pattern in which delayed re-quizzing has the greatest benefit for students who struggled the most in General Chemistry I.

Both General Chemistry I and Introductory Biology I are large introductory courses that are taken in the first year of college, when the majority of attrition within science occurs (Seymour and Hewitt, 1997; Chang *et al.*, 2008). Given the correlation in scores between these two courses, we should try to have tools for students to improve their performance during these first-year courses that could improve their persistence. Hence, because of our preliminary observation that delayed re-quizzing may help lower-achieving students (based on previous STEM course work) more than their higher-achieving peers, instructors might want to encourage delayed re-quizzing.

Intriguing is the lack of evidence for a similar interaction with incoming biology knowledge (i.e., pretest score). Why would delayed re-quizzing have differing effects based on previous achievement in STEM course work but not previous biology knowledge? While this finding certainly needs to be studied further, we provide one potential explanation. It is well established that lower-achieving students tend to report lower self-regulated learning skills (Zimmerman and Pons, 1986; Pintrich and de Groot, 1990; Kitsantas, 2002; Lopez *et al.*, 2013; Sebesta and Bray Speth, 2017), which includes tasks such as selecting effective study strategies. This is supported by our own data, in which chemistry exam performance predicted participation in re-quizzing a semester later in a different course. Given this, we expect lower-achieving students are more likely drawing from a pool of less effective study strategies, while higher-achieving students are more likely drawing from more effective strategies. We speculate that reopening the quizzes as a study tool for the exams might make self-testing more easily accessible to these lower-achieving students. It might be that lower-achieving students who took advantage of delayed re-quizzing may have incorporated this effective strategy into their repertoire of otherwise less-effective strategies, whereas higher-achieving students may have incorporated self-testing into their repertoire of otherwise more-effective strategies. In the future, it would be interesting to directly correlate the effectiveness of re-quizzing with study habits.

Implications for Instruction

This study reports that delayed no-stakes re-quizzing can be an effective learning tool for students, even when the quiz questions and exam questions are not identical. This suggests that the practice of reopening online quizzes that contain pools of questions (or rereleasing paper copies of multiple versions of quizzes) as study tools for exams, which many of us already do or could easily implement, can indeed benefit student learning. Further, to give timely feedback to our large class, our quizzes were solely closed-response questions (multiple choice, matching, etc.), yet they still generated a testing effect on our cumulative assessments. Because answering any high-level question requires some lower-level retrieval, regardless of the cognitive level of student-learning goals, retrieval practice (such as delayed re-quizzing) may be important for all instructors to consider.

Future Directions

Though we tried to make self-testing accessible, it was still underappreciated by students relative to more passive study strategies. For example, about half of our students reported that rewatching lecture recordings was “extremely helpful” for their learning, while less than a quarter reported that quizzing was. This fits with previous studies, which have similarly found that students have “illusions of competence” from more passive strategies, such as rereading, and tend not to view self-testing as a very effective learning strategy (Carrier, 2003; Karpicke *et al.*, 2009; Karpicke and Blunt, 2011; Kornell and Son, 2009). Despite this, more of our students chose to participate to some extent in re-quizzing (77% of students retook at least one quiz) than we expected based on published surveys (Karpicke *et al.*, 2009; Hartwig and Dunlosky, 2012). However, on the basis of published reports, we suspect that many of our students are using self-testing solely as a way to assess their preparedness, rather than as a learning tool (Kornell and Bjork, 2007; Karpicke *et al.*, 2009). That shift will be considerably harder to achieve. Some research suggests that, with guidance (including in-class demonstrations on the value of retrieval practice; Einstein *et al.*, 2012), students can begin to make more accurate judgments about the effectiveness of self-testing toward learning (Tullis *et al.*, 2013). This guidance toward helping students perceive the benefits of evidence-based learning strategies, though time-consuming, has the potential to benefit students throughout their undergraduate education.

ACKNOWLEDGMENTS

We thank Anton Weisstein, Barbara Kunkel, John Majors, and Kathleen Weston-Hafer for their consistent efforts to develop and revise the course pre/posttest and their help in revising quiz question pools. This research was supported in part by an internal grant, “Transformational Initiative for Educators in STEM,” which aims to foster the adoption of evidence-based teaching practices in science classrooms at Washington University in St. Louis.

REFERENCES

Bjork, E. L., Little, J. L., & Storm, B. C. (2014). Multiple-choice testing as a desirable difficulty in the classroom. *Journal of Applied Research in Memory and Cognition*, *3*(3), 165–170.

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

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In Healy, A., Kosslyn, S., & Shiffrin, R. (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (2nd ed., pp. 35–67). Hillsdale, NJ: Erlbaum.

Brame, C. J., & Biel, R. (2015). Test-enhanced learning: The potential for testing to promote greater learning in undergraduate science courses. *CBE—Life Sciences Education*, *14*(2), es4.

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, *36*(3), 604–616.

Carnegie, J. (2015). Use of feedback-oriented online exercises to help physiology students construct well-organized answers to short-answer questions. *CBE—Life Sciences Education*, *14*(3), ar25.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1563–1569.

Carrier, L. M. (2003). College students’ choices of study strategies. *Perceptual and Motor Skills*, *96*(1), 54–56.

Chang, M. J., Cerna, O., Han, J., & Sàenz, V. (2008). The contradictory roles of institutional status in retaining underrepresented minorities in biomedical and behavioral science majors. *Review of Higher Education*, *31*(4), 433–464.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159.

Eddy, S. L., & Hogan, K. A. (2014). Getting under the hood: How and for whom does increasing course structure work? *CBE—Life Sciences Education*, *13*(3), 453–468.

Einstein, G. O., Mullet, H. G., & Harrison, T. L. (2012). The testing effect: Illustrating a fundamental concept and changing study strategies. *Teaching of Psychology*, *39*(3), 190–193.

Freeman, S., Haak, D., & Wenderoth, M. P. (2011). Increased course structure improves performance in introductory biology. *CBE—Life Sciences Education*, *10*(2), 175–186.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, *48*(1), 26–34.

Haak, D. C., HilleRisLambers, J., Pitre, E., & Freeman, S. (2011). Increased structure and active learning reduce the achievement gap in introductory biology. *Science*, *332*(6034), 1213–1216.

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review*, *19*(1), 126–134.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In John, O., Robins, R., & Pervin, L. (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 114–158). New York: Guilford.

Kang, S. H. K., Mcdermott, K. B., Roediger, H. L., & Kang, S. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*(45), 528–558.

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*(6018), 772–775.

Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory*, *17*(4), 471–479.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*(5865), 966–968.

Kitsantas, A. (2002). Test preparation and performance: A self-regulatory analysis. *Journal of Experimental Education*, *70*(2), 101–113.

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, *14*(2), 219–224.

Kornell, N., & Son, L. K. (2009). Learners’ choices and beliefs about self-testing. *Memory*, *17*(5), 493–501.

Larsen, D. P., Butler, A. C., & Roediger, H. L., III (2009). Repeated testing improves long-term retention relative to repeated study: A randomised controlled trial. *Medical Education*, *43*(12), 1174–1181.

Leeming, F. C. (2002). The Exam-A-Day procedure improves performance in psychology classes. *Teaching of Psychology*, *29*(3), 210–212.

Lopez, E. J., Nandagopal, K., Shavelson, R. J., Szu, E., & Penn, J. (2013). Self-regulated learning study strategies and academic performance in undergraduate organic chemistry: An investigation examining ethnically diverse students. *Journal of Research in Science Teaching*, *50*(6), 660–676.

Lyle, K. B., & Crawford, N. A. (2011). Retrieving essential material at the end of lectures improves performance on statistics exams. *Teaching of Psychology*, *38*(2), 94–97.

McDaniel, M. A., & Little, J. (2019). Multiple-choice and short-answer quizzing on equal footing in the classroom: Potential indirect effects of testing. In Dunlowsky, J., & Rawson, K. A. (Eds.), *The Cambridge handbook of cognition and education* (pp. 480–499). Cambridge, UK: Cambridge University Press.

McDaniel, M. A., & Masson, M. E. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(2), 371–385.

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., Mcdermott, K. B., & Roediger, H. L. (2013). Quizzing in middle-school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology*, *27*(3), 360–372.

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

McDermott, K. B., Agarwal, P. K., D’Antonio, L., Roediger, H. L., & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied*, *20*(1), 3–21.

Messineo, L., Gentile, M., & Allegra, M. (2015). Test-enhanced learning: Analysis of an experience with undergraduate nursing students. *BMC Medical Education*, *15*, 182.

Orr, R., & Foster, S. (2013). Increasing student success using online quizzing in introductory (majors) biology. *CBE—Life Sciences Education*, *12*(3), 509–514.

Pintrich, P. R., & de Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, *82*(1), 33–40.

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, *330*(6002), 335.

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rickard, T. C., & Pan, S. C. (2018). A dual memory theory of the testing effect. *Psychonomic Bulletin & Review*, *25*(3), 847–869.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning. *Psychological Science*, *17*(3), 249–255.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463.

Sebesta, A. J., & Bray Speth, E. (2017). How should I study for the exam? Self-regulated learning strategies and achievement in introductory biology. *CBE—Life Sciences Education*, *16*(2), ar30.

Seymour, E., & Hewitt, N. M. (1997). *Talking about leaving : Why undergraduates leave the sciences*. Boulder, CO: Westview.

Smith, M. A., & Karpicke, J. D. (2014). Retrieval practice with short-answer, multiple-choice, and hybrid tests. *Memory*, *22*(7), 784–802.

Storm, B. C., Friedman, M. C., Murayama, K., & Bjork, R. A. (2014). On the transfer of prior tests or study events to subsequent study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(1), 115–124.

Theobald, R., & Freeman, S. (2014). Is it the intervention or the students? Using linear regression to control for student characteristics in undergraduate STEM education research. *CBE—Life Sciences Education*, *13*(1), 41–48.

Thomas, R. C., & McDaniel, M. A. (2013). Testing and feedback effects on front-end control over later retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 437–450.

Thorndike, E. (1914). *The psychology of learning*. New York: Teachers' College, Columbia University.

Trumbo, M. C., Leiting, K. A., McDaniel, M. A., & Hodge, G. K. (2016). Effects of reinforcement on test-enhanced learning in a large, diverse introductory college psychology course. *Journal of Experimental Psychology: Applied*, *22*(2), 148–160.

Tullis, J. G., Finley, J. R., & Benjamin, A. S. (2013). Metacognition of the testing effect: Guiding learners to predict the benefits of retrieval. *Memory & Cognition*, *41*(3), 429–442.

Wooldridge, C. L., Bugg, J. M., McDaniel, M. A., & Liu, Y. (2014). The testing effect with authentic educational materials: A cautionary note. *Journal of Applied Research in Memory and Cognition*, *3*(3), 214–221.

Zimmerman, B. J., & Pons, M. M. (1986). Development of a structured interview for assessing student use of self-regulated learning strategies. *American Educational Research Journal*, *23*(4), 614–628.