Received: 6 May 2019 | Revised: 29 July 2019 | Accepted: 6 September 2019
DOI: 10.1002/acp.3605

RESEARCH ARTICLE

WILEY

Frequent mastery testing with second-chance exams leads to enhanced student learning in undergraduate engineering

Jason W. Morphew[1] | Mariana Silva[2] | Geoffrey Herman[2] | Matthew West[3]

[1]School of Engineering Education, Purdue University, Indiana

[2]Department of Computer Science, University of Illinois at Urbana–Champaign, Champaign, Illinois

[3]Department of Mechanical Science and Engineering, University of Illinois at Urbana–Champaign, Champaign, Illinois

Correspondence
Jason W. Morphew, School of Engineering Education, Purdue University, West Lafayette, IN.
Email: jmorphew@purdue.edu

Funding information
National Science Foundation, Grant/Award Numbers: CMMI-1150490, DUE-1347722

Summary

Laboratory studies have routinely demonstrated that testing often leads to greater learning and retention than repeated studying. In the classroom, this effect has been replicated with memory and application tasks. However, studies of classrooms involving mathematical problem solving are sparse and have had mixed results. This paper presents the results of a quasi-experimental study in an undergraduate science, technology, engineering, and mathematics course that investigated more frequent testing that incorporated aspects of mastery testing and second-chance testing. Students in the frequent testing cohort scored seven percentage points higher and earned twice the number of As and half the number of failing grades. The advantage of frequent second-chance mastery testing was found for both multiple-choice and free-response questions and remained after controlling for differences in student ability. Women and underrepresented minority students benefited from the altered testing environment to the same extent as the general population.

KEYWORDS

engineering, second-chance testing, STEM, test-potentiated learning

1 | INTRODUCTION

Research and reform efforts to improve science, technology, engineering, and mathematics (STEM) education have largely focused on eschewing the traditional lecture in favor of active learning (Freeman et al., 2014; Henderson, Beach, & Finkelstein, 2011). However, comparatively little attention has been paid toward transforming the traditional assessment paradigm of a few midterm examinations and a final examination coupled with traditional lectures. This lack of attention may be due to examinations being viewed as measurements of current learning rather than as mechanisms to enhance learning (Hartwig & Dunlosky, 2012). Laboratory studies in cognitive science have robustly demonstrated that learning and retention of that knowledge can be enhanced through testing that incorporates feedback (Roediger & Butler, 2011), increased use of formative assessment (Clark, 2012), and distributed practice (Carpenter, Cepeda, Rohrer, Kang, & Pashler, 2012). Efforts to translate these laboratory studies into classroom have found testing to be beneficial for memory and application tasks (e.g., McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013); however, there has been limited research on the effect of testing in classrooms that teach mathematical problem solving (e.g., Downs, 2015; Hennig, Staatz, Bond, Leung, & Singleton, 2019).

The lack of attention on assessment practices suggests that we lack an understanding of a critical means for improving students' persistence in STEM. Although over a quarter of students entering a bachelor's degree program enroll in a STEM major at some point in their career, only half of those students leave having completed a STEM degree (Chen, 2013). Although several factors affect students' decisions to persist, grades in introductory courses are strong predictors of persistence within STEM majors (Cromley, Perez, & Kaplan, 2016; King, 2015; Rask, 2010) and have been found to mediate the effect of content knowledge on persistence (Dai & Cromley, 2014).

This paper presents the results of a quasi-experimental study involving a sophomore-level solid mechanics course investigating the effects of transitioning from two midterm exams to seven shorter exams spaced throughout the semester. However, simply increasing the frequency of assessments within a course does not guarantee that students will be tested more frequently on every concept covered in the course. If instructors simply construct shorter exams, it is likely that some, or all, of the concepts will be assessed on only once. Therefore, the exams incorporated second-chance and mastery aspects to ensure that all of the concepts would be assessed more frequently. Under this assessment structure, students could retake examinations to improve their scores, whereas the shorter formative exams allowed students to repeatedly attempt problems for reduced credit. Scores from identical, cumulative final exams were analyzed to address the following overarching research question: Does altering an assessment plan to include shorter and more frequent testing that incorporates second-chance testing improve student learning?

2 | LITERATURE REVIEW

2.1 | Testing effect

Engaging learners in testing has been shown to produce better long-term retention than restudying in clinical studies (Darley & Murdock, 1971; Roediger & Karpicke, 2006) as well as in secondary and university classrooms (e.g., McDaniel et al., 2013; McDermott, Agarwal, D'Antonio, Roediger, & McDaniel, 2014). For example, McDermott et al. (2014) utilized a within-subjects design with middle school students where course material was randomly assigned to be either tested, or restudied, or neither tested nor restudied. Students recalled facts at a higher rate for course material that was tested than for course material in the other conditions. Similar results have been found using online quizzes testing factual recall with undergraduate students in a psychology course (Johnson & Kiviniemi, 2009).

Much of the research concerning the testing effect has used posttest questions that are either identical or very similar to the questions used during the study phase. However, some studies have demonstrated the benefit of testing for rephrased questions (McDaniel, Anderson, Derbish, & Morrisette, 2007), for analogical problem solving (Peterson & Wissman, 2018), and for inferential and application questions (Butler, 2010; Thomas, Weywadt, Anderson, Martinez-Papponi, & McDaniel, 2018).

In memory tasks, testing appears to primarily benefit items correctly recalled during the study phase (Butler & Roediger, 2007; Karpicke & Roediger, 2007). However, other studies have found enhanced learning following testing for items that were answered unsuccessfully during initial testing when learners are provided with personalized feedback and can restudy the tested material (Kang, McDermott, & Roediger, 2007; Richland, Kao, & Kornell, 2008). In addition, some studies have found that testing enhances performance on new but untested material (Chan, 2010; Little, Bjork, Bjork, & Angello, 2012; Pan & Rickard, 2018; however, see Little, Storm, & Bjork, 2011; Wooldrige, Bugg, McDaniel, & Liu, 2014), suggesting that providing students with feedback and incentives to revisit the material following testing can potentiate future learning.

In the laboratory, much of the research concerning the testing effect has focused on memory tasks, whereas research in the classroom has utilized content focused on declarative memory tasks, such as word pairs in second language learning (Kang, Gollan, & Pashler, 2013), factual recall in psychology (McDaniel, Roediger, & McDermott, 2007), short answer questions in medical education (Larsen, Butler, & Roediger, 2009), recalling facts from a lecture (Butler & Roediger, 2007), and multiple-choice questions involving recalling or applying definitions in a middle school science course (McDaniel, Agarwal, Huesler, McDermott, & Roediger, 2011).

The benefit of testing on calculation-based problem-solving tasks, such as those found in introductory STEM courses, is less clear. Some researchers have asserted that testing effects are lessened as the complexity of the information increases (Hanham, Leahy, & Sweller, 2017; van Gog & Sweller, 2015). However, other researchers have documented testing effects for more complex tasks such as reading comprehension and inference tasks, learning spatial relationships, and constructing concept maps (e.g., Johnson & Mayer, 2009; Karpicke & Aue, 2015).

To our knowledge, only a few studies have examined the benefits of testing in mathematical problem-solving contexts. Most of these studies have compared engaging novices in either testing or studying worked examples. For these studies, students who engaged in repeated studying of worked examples generally outperformed those who studied an example and then completed practice problems (Leahy, Hanham, & Sweller, 2015; van Gog & Kester, 2012; van Gog, Kester, & Paas, 2011). For example, van Gog et al. (2015) compared testing versus restudy in students engaged in learning problem solving from worked examples across four experiments and found no advantage for testing over repeated studying for problem-solving tasks involving electrical circuits or probability distributions. Conversely, Hennig et al. (2019) conducted a quasi-experimental study comparing two cohorts of pharmacy students enrolled in pharmacokinetics and pharmacodynamics courses. They found that students who were quizzed throughout the semester earned higher grades in the course and on the final exams than students who were not quizzed.

2.2 | Frequency of assessments

Although testing is an effective method for long-term learning in many cases, students do not always take advantage of the benefits of testing when studying. Left to their own devices, students typically select passive study strategies that focus on encoding, such as rereading, reviewing notes, or rewatching lectures (Karpicke, Butler, & Roediger, 2009). In addition, students, and especially lower performing students, often select inefficient study strategies such as massed practice (i.e., cramming). Although massed practice often facilitates short-term performance, it has a detrimental effect on long-term retention and creates false perceptions of mastery (Brown, Roediger, & McDaniel, 2014; Butler, 2010; Rohrer, Taylor, Pashler, Cepeda, & Wixted, 2005; Schmidt & Bjork, 1992).

The discrepancy between research in cognitive science on the benefits of testing and student behavior suggests that instructors should engage students in more frequent testing. However, the effect that increased testing has in classroom contexts has been mixed. Although laboratory studies have generally found that increasing the number of tests often leads to greater retention (e.g., Vaughn & Rawson, 2011), the effect of increasing the number of assessments in classroom contexts is less clear. A meta-analysis by Bangert-Drowns, Kulik, and Kulik (1991) found large learning gains for testing over no testing, but that increasing the frequency of assessment beyond three assessments resulted in smaller learning gains. More recently, Foss and Pirozzolo (2017) found that across four semesters, students assessed more frequently scored higher on a final exam with a small effect size (d = 0.16). However, the difference between the conditions was only significant for one of the semesters. In addition, other studies have not found greater student outcomes from increasing the frequency of course assessments. For example, in a quasi-experimental study, Downs (2015) increased the number of assessments from four midterm exams to four midterm exams plus 14 quizzes but did not find a benefit on either midterm or final exam scores. In another study, Deck (1998) randomly assigned students in a marketing course to complete either weekly or monthly examinations of course material. Students who completed weekly exams had higher exam averages than those completing monthly exams. However, no differences on the cumulative final exam were found.

Historically, there has been an interest in the effect of assessment frequency on learning within mathematical problem-solving courses (see Bangert-Drowns et al., 1991, for a review). More recently, the few studies investigating course assessment frequency have found mixed results. Stephens (1977) found no difference among test scores between students tested twice, four times, or weekly. Similarly, Dineen, Taylor, and Stephens (1989) found no difference on the final exam for high school math students randomly assigned to complete either daily or weekly quizzes. Additionally, Ward (1984) found that increasing the number of assessments from two per semester to weekly actually lead to worse performance on the final exam.

However, other studies have found an advantage for more frequent assessment schedules. Pikunas and Mazzota (1965) found that giving high school students enrolled in a Chemistry II course weekly tests led to higher scores on midterm exams as compared with restudying. Similarly, Townsend and Wheatley (1975) found that introductory calculus students who were quizzed daily had higher scores on an aptitude test designed for the study than those students who completed only one midterm but not higher than students completing three midterms or quizzed every fourth or fifth class.

2.3 | Mastery and second-chance testing

A productive strategy for encouraging effective long-term learning involves taking challenging tests for which immediate and productive feedback is given (Kang et al., 2007; Kornell, Hays, & Bjork, 2009; Roediger & Butler, 2011). In addition, providing students with the opportunity and motivation to restudy the material on which they were tested appears to lead to test-potentiated learning, a term used to describe the beneficial effects of studying following testing (Chan, Manley, Davis, & Szpunar, 2018; Pyc & Rawson, 2010; Rawson & Dunlosky, 2012). One potential strategy to encourage students to reflect on feedback and restudy the tested material is to shift toward a more formative assessment structure.

Summative assessments are often formal, “one-shot” assessments used to measure a student's current ability and assign grades. In contrast, formative assessments are often repeatable and used to provide ongoing feedback to the learner. Although formative assessments are generally thought to be more beneficial for future learning, summative assessments can also be used for formative purposes (Bell & Cowie, 2000). One example of a summative assessment being used for formative purposes is mastery testing. Assessments in the mastery testing framework are used to evaluate student progress by assigning a summative score and also to inform students about their progress and to incentivize corrective actions by allowing students to retake all or part of the exam until a mastery level is reached. Previous studies of mastery testing have found learning gains, particularly for lower ability students (e.g., Kulik & Kulik, 1987).

A simplified form of mastery testing is second-chance testing, which allows students to retake all or part of an exam at a later date. Such an approach recognizes that traditional summative assessments that allow for a single attempt measure a student's metacognitive ability to recognize when they have sufficiently prepared for an exam, as well as their ability or willingness to learn (Nelson, 1996). Previous studies of second-chance testing have found that a majority of students report greater positive affect and lower test anxiety (Diegelman-Parente, 2011; William, House, & Boyd, 1984). In addition, and perhaps unsurprisingly, many students who retake an exam earn higher exam scores on the retake (Juhler, Rech, From, & Brogan, 1998; Roszkowski & Speat, 2016). Although preliminary evidence indicates positive outcomes for second-chance testing, prior work is inconclusive about its impacts on retention of the course material as measured by final exam scores (Bangert-Drowns et al., 1991; Juhler et al., 1998). In addition, little is known about how best to employ second-chance and mastery testing in large introductory STEM courses without overwhelming instructors and teaching assistants.

3 | METHODS

The study described in this paper focuses on the effects of increasing the frequency with which course assessments are administered as well as changing the nature of the course exams from summative assessment to a hybrid approach that incorporates aspects of formative assessment. Because the primary goal of this research study is to explore the effects of changing the frequency and nature of assessments over the course of a semester, we employed a quasi-experimental design that prioritized the ecological validity of the study. As such, we compared consecutive semesters of a course taught by the same instructor with the same course content that differed in the frequency and nature of the assessments.

The sequential study design has the added benefit of avoiding ethical dilemmas arising from randomly withholding the treatment from students, potential diffusion of treatment resulting from students studying together or transferring between sections, and logistical challenges of requiring a single instructor to run two different versions of the course in parallel. The lecture and discussion section activities were also held constant, providing additional control. The homework assignments were kept mostly the same with some small variations, which are described in the section on course and assessment design.

To answer the research question, this study compared student performance on an identical final exam before and after the change in assessment schedule, using statistical tests to control for student ability and prior learning. Although we could not use random assignment for the treatments, thus limiting the causal claims that we can make, the ecological validity, large sample size, and similarity of the students provide a sufficiently controlled and well-powered study to conclude that the findings from the study may be considered robust. The research design and procedures were approved by the Institutional Review Board at Midwestern University.

3.1 | Participants

A total of 480 students enrolled in an introductory solid mechanics course at a large Midwestern university across two semesters were used for this study. Two hundred forty-eight students were enrolled in the semester with the traditional exam schedule, whereas 232 students were enrolled in the semester utilizing more frequent testing. Demographic data for the courses are provided in Table 1.

3.2 | Course and assessment design

The introductory solid mechanics course investigates the relationship between internal stresses and deformations produced by external forces acting on deformable bodies, as well as design principles based on mechanics of solids. In this course, students attended three 50-min lectures taught by the course instructor (second author) and one 50-min discussion section led by teaching assistants each week. During discussion sections, students worked on real-world engineering problems, open-ended design problems, and hands-on activities.

3.2.1 | Assessments

Students enrolled in the course during the traditional assessment schedule semester completed two midterm exams consisting of 10–13 multiple-choice questions and two or three free-response questions. Questions asked students to apply equations and theory to solve problems related to design using calculation (Figure 1). Multiple-choice questions were graded as either correct or incorrect with no partial credit, whereas free-response questions required students to show their work and could earn partial credit if they demonstrated correct conceptual understanding but made mathematical errors. Students received correctness feedback 1 week after the exams.

Students enrolled in the course during the more frequent assessment schedule completed seven computer-based exams, as well as five short in-lecture free-response quizzes. The computer-based exams were administered outside of class through a computer-based testing facility (Zilles, West, Mussulman, & Bretl, 2018). These exams consisted of three to five questions that asked students to apply equations and theory to solve mathematical engineering problems (Figure 2). In this format, students receive immediate correctness feedback and can go back and attempt the same question again for reduced credit, if time allows. After the exam, students can access the correct answers and then study and attempt to retry a different exam covering the same content the following week. The in-lecture, free-response quizzes consisted of one real-world application problem solved during the discussion section that asked students to apply the material covered in the lecture to solve a mathematical problem related to design. Students received correctness feedback one week after the quizzes.

3.2.2 | Homework

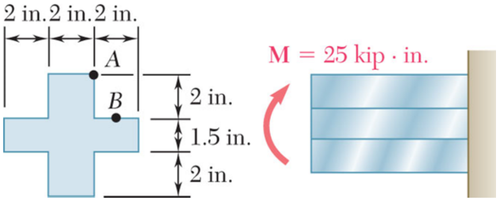
Homework was assigned in both a computer-based format and a paper-based format. In both semesters, the computer-based homework engaged students in solving engineering problems where they apply the equations and theories covered in lecture to solve problems related to design and failure analysis. For the traditional assessment group, the paper-based homework was assigned weekly and consisted of a real-world application problem that was solved using the standard

TABLE 1 Demographic and ability information by semester

Variable	Traditional exam schedule	More frequent exam schedule
Student gender	20% female and 80% male	21% female and 79% male
Student ethnicity	11% Asian American, 58% Caucasian, 20% International, and 10% URM	15% Asian American, 47% Caucasian, 28% International, and 9% URM
Average student grade in prerequisite course	2.98, 95% CI [2.87, 3.09]	3.19, 95% CI [3.06, 3.32]
Average ACT math score	31.8, 95% CI [31.3, 32.2]	32.9, 95% CI [32.4, 33.3]

Abbreviations: ACT, American College Testing; CI, confidence interval; URM, underrepresented minority.

(a) 8/1. (1 point) The cross section below has moment of inertia with respect to the centroidal axis given by $I = 28.85$ in^4. Determine the magnitude of the normal stress at point B.



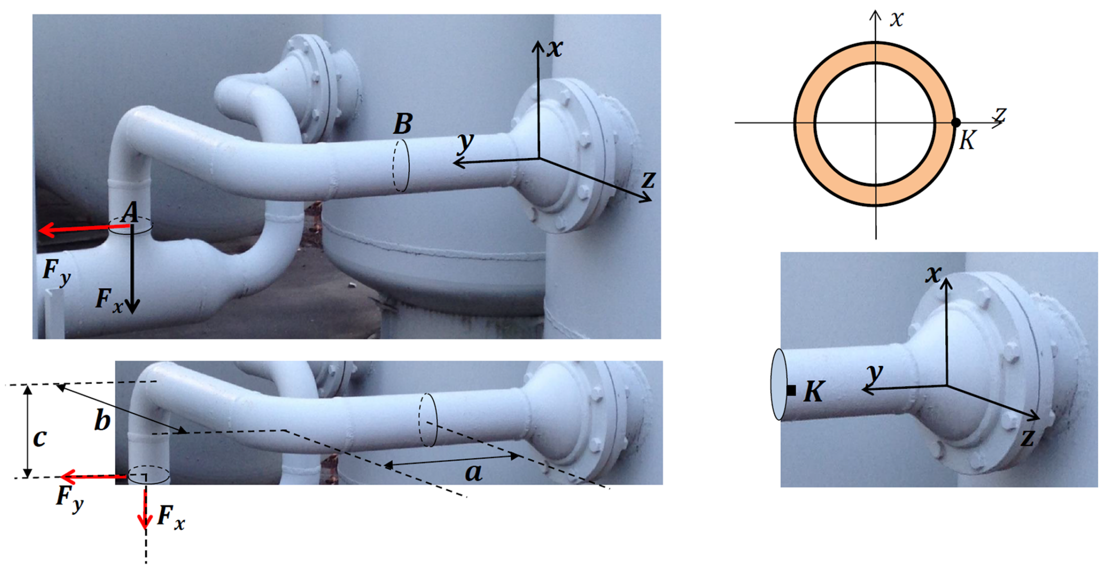
A. 1.3 ksi

B. 2.38 ksi

C. ★ 0.65 ksi

D. zero

(b) 14. (4 points) For the pipe system below, let us assume that the internal forces at cross-section A can be represented by an axial force $F_x = 3$ kips (e.g. the weight of the pipe and fluid) and a shear force in the y-direction $F_y = 2$ kips (we simplify this problem by neglecting all the other internal forces and moments at A). The gage pressure is given by $p = 2$ ksi. Use the dimensions $a = 3$ ft, $b = 2$ ft and $c = 1$ ft. The pipe has outside diameter $d_o = 8$ in, internal diameter $d_i = 7.5$ in and is made of a material with yielding strength $\sigma_Y = 36$ ksi. Assuming that point K is the critical point for this pipe, determine if the pipe fails according to Tresca Criteria.



Cross-section properties:

Moment of inertia: $I = \frac{\pi(d_o^4 - d_i^4)}{64} = 45.7$ in^4

Area: $A = \frac{\pi(d_o^2 - d_i^2)}{4} = 6.09$ in^2

First moment of half-circle: $Q = \frac{(d_o^3 - d_i^3)}{12} = 7.51$ in^3

FIGURE 1 Example final exam questions: (a) multiple-choice and (b) free-response questions [Colour figure can be viewed at wileyonlinelibrary.com]

given–find–solution engineering format (Sobek, Cundy, & Briggeman, 2004). Students received correctness feedback and were graded on their problem-solving accuracy. For the frequent assessment group, the paper-based homework was amended to reduce the amount of time students engaged in mathematical problem solving and give students practice with technical writing and group collaboration skills. At four points during the semester, groups of students were asked to apply the equations and theories covered in lecture to iteratively evaluate one aspect of a mechanical design problem. The revised paper-based homework engaged students in problem-solving tasks similar to

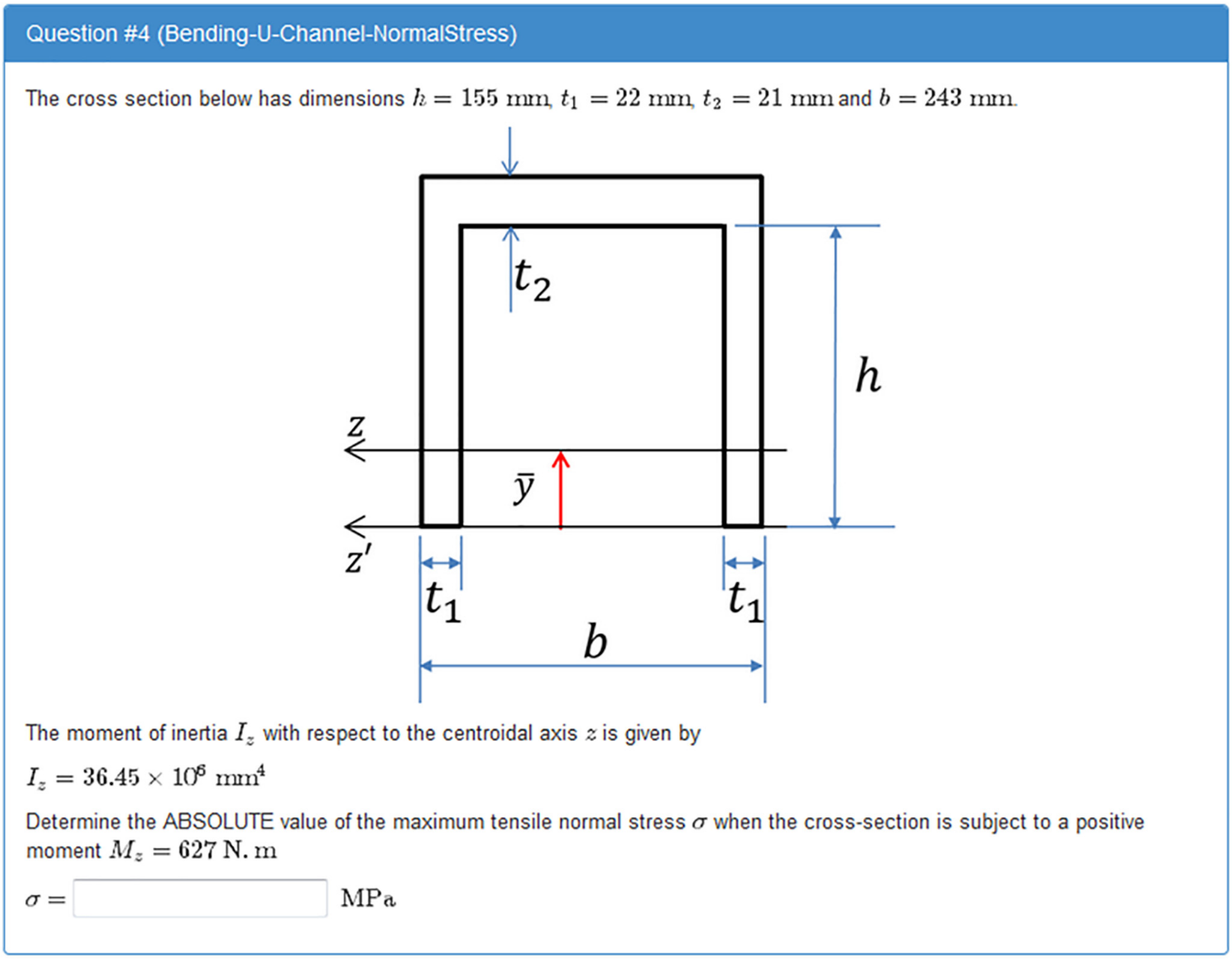


FIGURE 2 Example of a computer-based quiz question [Colour figure can be viewed at wileyonlinelibrary.com]

the previous semester; however, students spent less time completing homework and were asked to solve fewer problems (four compared with nine). Students received correctness feedback; however, the grading emphasized written communication skills. In addition, students no longer had the opportunity to practice solving problems by hand following the given–find–solution format expected on the free-response section of the final exam.

3.2.3 | Final exam

The final exam was identical across both semesters and was composed of 13 multiple-choice questions and three free-response questions in the given–find–solution format. The questions were new to students in both semesters; however, they covered the same concepts as the exams and homeworks completed by the students in both semesters. The multiple-choice section engaged students in solving mathematical problems related to design and were graded as either correct or incorrect with no partial credit. The free-response questions engaged students in solving problems related to design using calculation and required students to show their work. On this section, students could earn partial credit if they demonstrated correct conceptual understanding but simply made mathematical errors.

Given the changes in students' exposure to given–find–solution-type problems during the semester, we evaluated the effect of the change in assessment schedule on the overall exam score and for each section of the exam separately.

3.3 | Procedure

To investigate the overarching research question of whether altering an assessment schedule improves student performance on the final exam, four subquestions were examined. First, to determine whether the average performance on the final exam differed between the semesters, an independent-samples t test was conducted with final exam score as the response variable and semester as the between-subjects variable. The distribution of the exam scores was generally not normally distributed; however, t tests are typically considered robust to deviations from normality with large sample sizes. To be conservative, nonparametric Kruskal–Wallis tests were also conducted when the scores were not normally distributed. The conclusions reached using the more conservative approach are similar to the results from the t tests; therefore, only the t test results are reported.

Second, to determine whether differences in final exam scores were the result of differences in student ability or demographic differences between the semesters, four regression models were fit with final exam score as the response variable and semester as the between-subjects variable. Due to the skewed distribution and the presence of a ceiling effect, we fit beta regression models to control for demographic factors and prior ability. Because not all information was available for all students, the models were fit in a hierarchical manner, including variables with less missing information first. In Model 1, demographic variables available for all students were included in the model—gender, international (whether an individual was a domestic or international student), and underrepresented minority status (URM). In Model 2, a measure of student ability—grade in the prerequisite course (introductory statics)—was added because this information was available for most students. In Model 3, a second measure of student ability—American College Testing (ACT) math score—was included. Models 4 and 5 investigate whether increased testing differentially facilitated learning for students of differing ability levels by including interactions between semester and measures of prior ability.

Third, to examine whether numeric changes in the final exam average resulted in qualitative changes in letter grades, chi-squared tests of independence and a logistic regression were conducted. Fourth, to determine whether differences in performance on the final exam were due to question type, independent-samples *t* tests, beta regressions, chi-squared tests of independence, and logistic regressions were conducted on the multiple-choice and free-response sections.

4 | RESULTS

4.1 | Effect on final exam performance

Students enrolled in the more frequent exam semester scored about seven percentage points higher on the final exam compared with students in the traditional assessment semester (Table 2). The grades on the final exam were not normally distributed; however, homogeneity of variance could be assumed, and the distributions were similarly negatively skewed. The difference was significant, $t(478) = 5.58$, $p < .001$, with a medium effect size, $d = 0.51$, 95% confidence interval, CI [0.33, 0.69], roughly equivalent to two thirds of a letter grade.

TABLE 2 Means and standard deviations for final exam scores

Exam Score	Traditional assessment schedule	Frequent exam schedule
Final exam score	76.7 (14.0)	83.6 (13.0)
Multiple-choice average	79.9 (14.5)	83.4 (14.1)
Free-response average	72.5 (17.0)	83.8 (15.3)

4.2 | Final exam performance controlling for demographics and differences in ability

Although the students enrolled in the introductory solid mechanics course had similar demographic profiles across both semesters, it is possible that there was a difference in student ability between the semesters. Most students (traditional, $n = 233$; frequent, $n = 205$) completed the prerequisite course on campus and had their final prerequisite course grade available, whereas fewer students had their ACT math score available (traditional, $n = 187$; frequent, $n = 149$). Although any difference in grade in the prerequisite course might merely represent differences in course policies or strictness in grading between the instructors in the prerequisite course, we analyzed the distribution of letter grades from the prerequisite course and noted that students enrolled in the semester with more frequent testing were more likely to earn an A and less likely to earn a C in the prerequisite course. A chi-squared test of independence indicated that student grades in the prerequisite course differed between semesters, $\chi^2(4) = 12.4$, $p < .05$. Follow-up tests indicated that students enrolled in the semester with more frequent testing were 1.5 times more likely to earn an A, 95% CI [1.2, 1.9], and 1.3 times less likely to earn a C, 95% CI [0.9, 1.9], in the prerequisite course. In addition, a Kruskal–Wallis test on ACT math scores indicated that students assessed with the more frequent exam schedule had higher ACT math scores, $\chi^2(1) = 11.16$, $p < .01$.

To control for demographic factors and student ability, we conducted a sequential regression analysis using a beta distribution using PROC GLIMMIX in SAS version 9.4. Because beta regression models values between 0 and 1, the percentage on the final exam was expressed as a decimal and was translated by subtracting .01 from all values to avoid any scores at the boundary (Smithson & Verkuilen, 2006). The conditional means were modeled with the logit link for the beta regression models, and to avoid issues of multicollinearity, ACT math scores were centered by subtracting the mean from the ACT math scores.

Pseudo-R^2 values were computed by computing the squared correlation coefficients between the linear predictors from the models and the link-transformed response variables (Ferrari & Cribari-Neto, 2004) to compare model fit. The results of the beta regressions are found in Table 3 (Models 1–3). The models indicate that students who were assessed with the more frequent exam schedule scored higher on the final exam even after controlling for demographics and differences in prior ability. Although Model 3 has the largest pseudo-R^2 value, the loss of subjects due to missing data is relatively large. For this reason, we opt to interpret Model 2 because this model balances the need for control with subject retention.

The estimated parameter indicates that the logit for a student enrolled in the more frequent testing semester is 1.4 times higher than the logit for a student in the traditional testing semester when controlling for student ability and demographics. In other words, students who engaged in more frequent testing outperformed students who engaged in the traditional exam schedule on average even after controlling for demographics and differences in prior ability.

TABLE 3 Association between total exam score and exam schedule controlling for demographic variables and student ability using beta regression

	Model 1		Model 2		Model 3		Model 4		Model 5	
Variable	*β* (*SE*)	*p*	*β* (*SE*)	*p*	*β* (*SE*)	*p*	*β* (*SE*)	*p*	*β* (*SE*)	*p*
Gender	0.190 (0.08)	.03	0.127 (0.08)	.09	0.090 (0.08)	.28	0.128 (0.08)	.09	0.093 (0.08)	.26
International	0.148 (0.09)	.08	0.066 (0.08)	.39	0.103 (0.17)	.54	0.067 (0.08)	.38	0.096 (0.17)	.57
URM	−0.063 (0.12)	.59	−0.001 (0.11)	.99	0.031 (0.10)	.76	0.004 (0.11)	.97	0.032 (0.10)	.76
Prerequisite			0.517 (0.03)	<.001	0.518 (0.04)	<.001	0.531 (0.05)	<.001	0.519 (0.04)	<.001
ACT math					0.033 (0.01)	<.01			0.049 (0.02)	.02
Semester	0.420 (0.07)	<.001	0.330 (0.06)	<.001	0.334 (0.07)	<.001	0.334 (0.07)	<.001	0.331 (0.07)	<.001
Prerequisite × Semester							0.028 (0.07)	.68		
ACT Math × Semester									0.022 (0.02)	.39
Pseudo-R^2	.07		.39		.47		.39		.47	
AIC	−679.55		−806.11		−626.32		−804.28		−625.13	
BIC	−654.51		−777.53		−596.33		−771.62		−591.38	
N	480		438		314		438		314	

Abbreviations: ACT, American College Testing; AIC, Akaike information criterion; BIC, Bayesian information criterion; URM, underrepresented minority.

4.3 | Effect on distribution of letter grades

To examine the practical effect of the increase in scores on student outcomes, the letter grade on the final exam for each student was calculated using a commonly applied grade scale (A ≥ 90%, 90% > B ≥ 80%, etc.). The difference between the semesters is visualized in Figure 3. An overall chi-squared test of independence indicated that the two grade distributions differed, $\chi^2(3) = 33.21$, $p < .001$. To examine this further, the frequency of A grades and failing grades on the final exam were examined. To determine whether differences in the letter grades on the final exam were due to differences in ability between the semesters rather than differences in assessment schedule, we conducted two logistic regressions with the frequency of A and failing grades as the response variables and course, gender, international, URM, and prerequisite course grade as the criterion variables. The results indicate that students who were assessed with the more frequent exam schedule were more likely to receive an A, $\chi^2(1) = 11.14$, $p < .001$, odds ratio [OR] = 2.4, 95% CI [1.4, 3.9], and less likely to

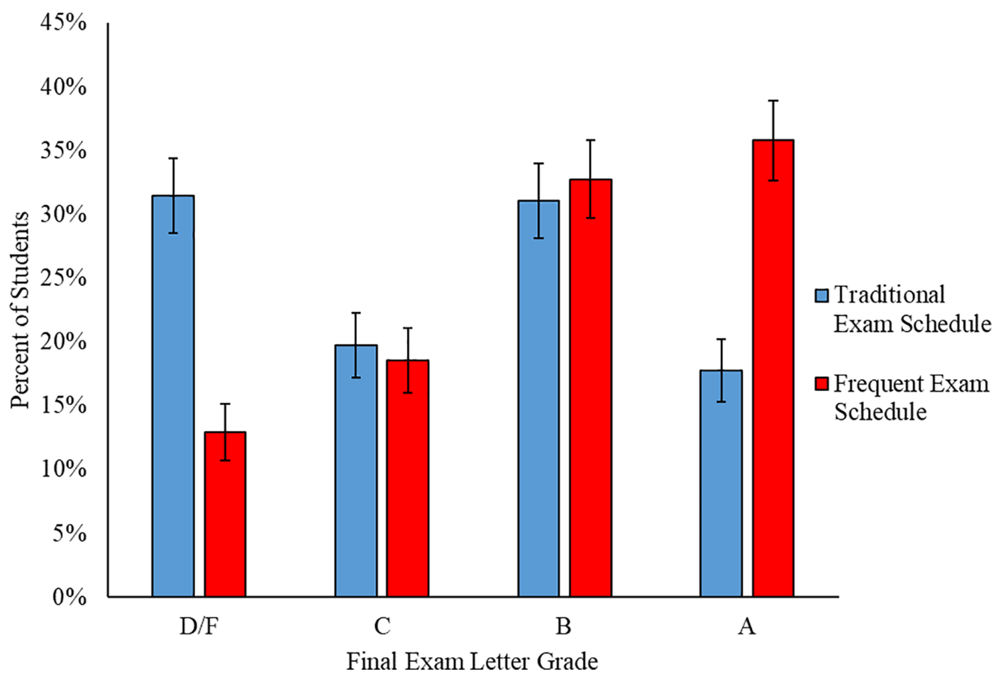


FIGURE 3 Histogram of student results on the identical final exam for the two semesters. Exam letter grades correspond to the following exam score ranges: A (90% or higher), B (80% to 90%), C (70% to 80%), and D/F (below 70%). Error bars show 95% confidence intervals [Colour figure can be viewed at wileyonlinelibrary.com]

receive a failing grade, $\chi^2(1) = 18.66$, $p < .001$, OR = 0.3, 95% CI [0.2, 0.5], on the final exam, when controlling for the grade in the prerequisite course.

4.4 | Effect on final exam score by subgroups

To examine whether the move to more frequent testing differentially affected particular subgroups of students, three analyses were conducted. To examine whether more frequent testing differentially affects students of different ability levels, two beta regression models with ability–semester interactions were conducted (Table 3; Models 4 and 5). In both models, the interactions were not significant, indicating that the intervention appears to benefit students of all ability levels. The descriptive statistics for the final exam scores for each subgroup are found in Table 4 and visualized in Figures 4 and 5. To examine the effect of increased testing on underrepresented groups such as female and URM students, separate beta regressions were conducted for each group (see Tables 5 and 6). The regression coefficients were tested using equations (1) and (4) in Paternoster, Brame, Mazerolle, and Piquero (1998). The differences between the regression coefficients for each group were not significantly different, suggesting that more frequent testing benefitted both majority and underrepresented groups similarly.

4.5 | Effect on final exam scores by question type

Students assessed with the more frequent exam schedule scored higher on the multiple-choice questions by 3.5 percentage points, $t(478) = 2.71$, $p = .007$. A greater percentage of students under the more frequent exam schedule earned a perfect score on the multiple-choice section of the final exam (20.7% compared with 9.7%). Because this represents a potential ceiling effect, the frequency of perfect scores between the semesters was examined in addition to the frequency of A and failing grades for the multiple-choice section. To determine whether differences in the letter grades were due to differences in ability between the semesters, three logistic regressions were conducted with the frequency of perfect scores, A grades, and failing grades as the response variables, and course, gender, international, URM, and prerequisite course grade as the criterion variables. The results indicate that students who were assessed with the more frequent exam schedule were more likely to receive a perfect score, $\chi^2(1) = 6.25$, $p = .01$, OR = 2.1, 95% CI [1.2, 3.8], and less likely to receive a failing grade, $\chi^2(1) = 5.90$, $p = .015$, OR = 0.5, 95% CI [0.3, 0.9], on the multiple-choice section of the final exam, when controlling for demographics and grade in the prerequisite course. However, due to ceiling effects, students who were assessed with the more frequent exam schedule were not more likely to receive an A on the multiple-choice section, $\chi^2(1) = 1.59$, $p = .21$, when controlling for demographics and grade in the prerequisite course.

On the free-response questions, students assessed with the more frequent exam schedule scored higher on the free-response questions by 11.3 percentage points, $t(478) = 7.59$, $p < .001$, on the final exam. To determine whether differences in the letter grades were due to differences in ability between the semesters, two logistic regressions were conducted with the frequency of A's and failing grades as the response variables and course, gender, international, URM, and prerequisite course grade as the criterion variables. The results indicate that students who were assessed with the more frequent exam schedule were more likely to receive an A, $\chi^2(1) = 28.48$, $p < .001$, OR = 4.4, 95% CI [2.5, 7.5], and less likely to receive a failing grade, $\chi^2(1) = 33.83$, $p < .001$, OR = 0.2, 95% CI [0.1, 0.3], on the free-response section of the final exam when controlling for demographics and grade in the prerequisite course.

5 | DISCUSSION

In this study, we found that students enrolled in an introductory solid mechanics course scored higher on the final exam when the assessment schedule was modified from two large, "one-shot" midterm exams to seven shorter exams with second-chance mastery testing. Students who completed the more frequent testing schedule scored, on average, seven percentage points higher on an identical final exam than students who completed the traditional assessment schedule. More importantly, the percentage of students receiving a grade of A on the final exam was twice as large (36% vs. 18%), and the percentage of students receiving failing grades (Ds or Fs) was about two and a half times lower (13% vs. 31%) under the more frequent assessment schedule. This is an important threshold in many introductory courses because students do not earn course credit for earning either Ds or Fs and are less likely to persist in the major if they earn low grades in introductory courses (e.g., King, 2015). Women and underrepresented

TABLE 4 Means and standard deviations for final exam scores by underrepresented demographic groups

	Female		Male		URM		Non-URM	
Exam Score	Traditional schedule (n = 49)	Frequent schedule (n = 48)	Traditional schedule (n = 199)	Frequent schedule (n = 184)	Traditional schedule (n = 26)	Frequent schedule (n = 22)	Traditional schedule (n = 222)	Frequent schedule (n = 210)
Overall final exam	74.5 (12.8)	81.8 (11.5)	77.2 (14.2)	84.0 (13.4)	75.5 (9.8)	84.3 (10.4)	76.8 (14.4)	83.5 (13.3)
Multiple choice	77.4 (14.2)	79.7 (13.4)	80.4 (14.5)	84.4 (14.1)	78.1 (12.1)	83.2 (12.0)	80.1 (14.7)	83.4 (14.3)
Free response	70.8 (15.5)	84.6 (12.2)	72.9 (17.4)	83.5 (16.1)	72.2 (13.9)	85.6 (10.2)	72.6 (17.4)	83.6 (15.8)

Abbreviation: URM, underrepresented minority.

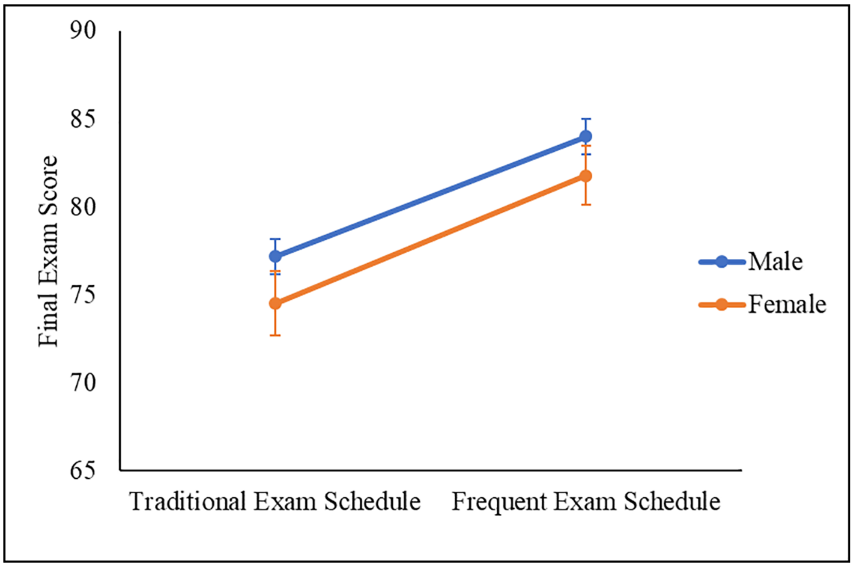


FIGURE 4 Scores on the final exam by semester for male and female students [Colour figure can be viewed at wileyonlinelibrary.com]

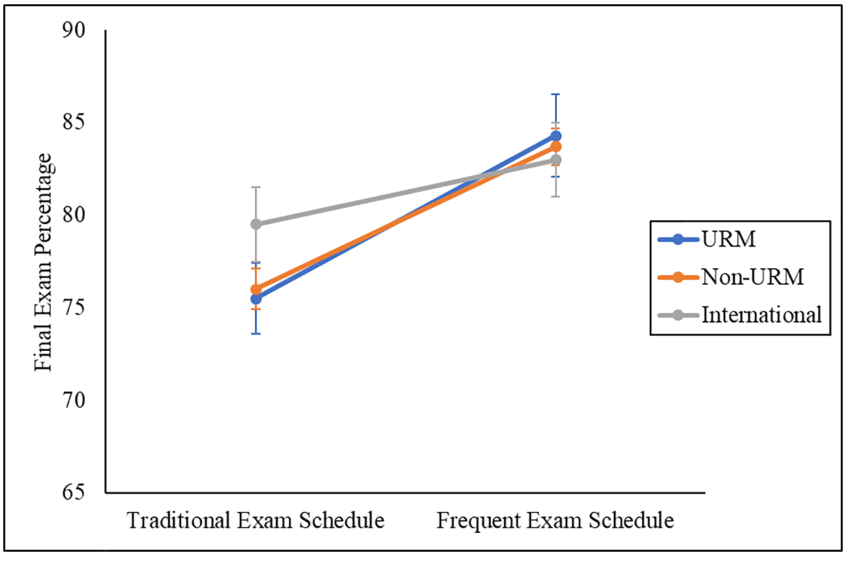


FIGURE 5 Scores on the final exam by semester for underrepresented minority (URM), international, and non-URM students [Colour figure can be viewed at wileyonlinelibrary.com]

minority students saw the same increases as the rest of the student population. Because the average scores of women were lower in this population, they experienced a greater benefit at the lower end of the distribution. International students saw smaller improvements in average final exam scores, partially because these students scored higher in the semester with less frequent testing.

Prior studies have found that assessment schedules incorporating more frequent exams often lead to better performance on those exams; however, the improved performance is not always observed on a comprehensive final exam (Deck, 1998; Downs, 2015; Juhler et al., 1998). The findings from this study suggest that the addition of second-chance testing delivered within a computer-based mastery

TABLE 5 Association between total exam score and exam schedule controlling for other demographic variables by gender

	Female (n = 97)		Male (n = 383)	
Variable	β (SE)	p	β (SE)	p
International	0.467 (0.16)	.005	0.054 (0.10)	.58
URM	−0.070 (0.24)	.77	−0.066 (0.13)	.61
Semester	0.379 (0.14)	.006	0.429 (0.08)	<.001
Test of the regression slopes for semester	$t(476) = 0.31$, $p = .75$			

Abbreviation: URM, underrepresented minority.

TABLE 6 Association between total exam score and exam schedule controlling for other demographic variables by URM status and *t* test of regression slope for semester variable compared with domestic non-URM students

	Non-URM (*n* = 319)		URM (*n* = 48)		International (*n* = 113)	
Variable	*β* (*SE*)	*p*	*β* (*SE*)	*p*	*β* (*SE*)	*p*
Gender	0.290 (0.10)	.004	0.288 (0.22)	.21	−0.087 (0.19)	.64
Semester	0.453 (0.08)	<.001	0.602 (0.18)	<.001	0.257 (0.16)	.11
Test of difference in regression slopes for semester			$t(363) = 0.75$, $p = .45$		$t(428) = 1.09$, $p = .28$	

Abbreviation: URM, underrepresented minority.

testing framework may provide added value for learning in addition to those found with test-enhanced learning.

Although students tend to use self-testing to measure current learning rather than to enhance learning, the use of testing as a study strategy is strongly correlated with grades (Hartwig & Dunlosky, 2012). However, little research has investigated individual differences in test-enhanced learning. The effect of ability on learning from testing is mixed, with some studies finding that testing was effective for high-performing students but not for middle- and low-performing students, whereas others found no differences in learning from using testing across ability groups (Rawson, Dunlosky, & Sciartelli, 2013). In the current study, both low- and high-performing students saw similar increases on the final exam as indicated by the lack of significant interactions in the regression models. Additionally, more frequent testing appears to benefit female and URM students similarly to male and domestic non-URM students, given the nonsignificant difference in the regression coefficients.

Coursework in engineering requires students to be able to construct well-developed and connected conceptual understandings, in addition to building the strong retrieval pathways necessary for memory tasks, such as problem categorization and procedural knowledge, in order to successfully solve problems typically found in introductory engineering courses. Much of the research concerning testing frequency has focused on memory tasks and has found that engaging in more frequent testing improves future performance by strengthening the representations in memory. The present investigation extends this work by examining the testing effect in an ecologically valid and relatively long-term application by comparing the effects of traditional, less frequent testing to more frequent testing in a semester-long engineering course.

Simply increasing the frequency of testing within a course does not guarantee that students will be tested more frequently on the concepts covered in the course. We implemented the second-chance and mastery aspects of the exams to ensure that students would be assessed more frequently on all concepts. It is likely that the combination of increased frequency of assessments with second-chance mastery testing provides additional benefits that explain why prior studies have not documented the same effect for increased testing frequency for mathematical problem-solving tasks. Previous work has found that allowing students to reattempt problems using mastery testing and second-chance exams resulted in better final exam performance (Bangert-Drowns et al., 1991; Juhler et al., 1998; Kulik & Kulik, 1987). Although we cannot disentangle the effects because we changed both aspects of the assessment schedule at the same time, we offer some theoretical reasons why the combination of increased frequency and second-chance testing may provide additional benefit for students' learning in order to inform the design of future studies.

The use of second-chance testing likely motivates students to engage in studying the exam material after receiving feedback on their performance. Evidence for this interpretation comes from the fact over 70% of the students elected to attempt the second try on each exam, including students who earned high (or even perfect) scores on their first attempt. The use of more frequent assessments with second chance testing may have led students to take advantage of test-potentiated learning more than under the traditional assessment schedule, allowing students to learn more effectively from studying after testing (Bjork & Storm, 2011; Chan et al., 2018; Wissman, Rawson, & Pyc, 2011). Test-potentiated learning appears to indirectly improve learning through enhanced metacognitive monitoring and self-regulated learning strategies, as well as through direct effects of testing on the encoding of new information (Fernandez & Jamet, 2017). Testing also encourages learners to shift toward more effective study strategies by giving students feedback on the effectiveness of their strategy use (Cho & Powers, 2019; Finley & Benjamin, 2012; Soderstrom & Bjork, 2014). In this study, test-potentiated learning may have stimulated deeper conceptual learning, or development of problem-solving skills, for students engaged in more frequent testing that has not been seen in previous studies.

An alternate explanation is that the combination of increased frequency and second-chance testing encouraged students to engage in more distributed practice. Left to their own devices, students often engage in massed practice (i.e., cramming) when studying for exams. A number of studies have found that up to two thirds of undergraduate college students report using cramming as a primary study strategy and over half report the tendency to study in one session immediately before a test (Blasiman, Dunlosky, & Rawson, 2017; Hartwig & Dunlosky, 2012). With a second-chance test a week later, students are incentivized to engage in a more distributed form of study, which may have led to the observed improvement on the cumulative final exam. In addition, the spacing of the second-chance exam from the original exam by a week is beneficial for long-term retention as prior research has found that the optimum spacing of studying is about 10–30% of the desired retention interval (Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008; Son & Simon, 2012). In other words, allowing students to retake exams disrupts the unproductive pattern of massed practice immediately before exams. A beneficial line of future research would be to investigate differences in the study strategies and schedules that students use when offered different permutations of increased frequency of assessments and second-chance testing.

Future work also should investigate which components of the new assessment schedule are beneficial (i.e., increased frequency vs. second-chance testing vs. mastery testing) or if the combination of all features is essential for observing learning gains. For example, future work should attempt to disentangle the effects of frequent testing from those of second-chance testing. Regardless of the exact causality, the combination of frequent testing, second-chance testing, and mastery testing appears to be a promising new assessment paradigm that could dramatically improve outcomes in STEM education and merits more research.

ACKNOWLEDGMENTS

This research was supported in part by the National Science Foundation under Grants DUE-1347722 and CMMI-1150490. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The data that support the findings of this study are subject to the Family Educational Rights and Privacy Act and are therefore not publicly available due to privacy and ethical restrictions. Anonymized data may be available from the corresponding author upon reasonable request; however, requests for data are subject to approval from the university's internal review board.

CONFLICT OF INTERESTS

The second author served as the course instructor during the study; however, they did not participate in the collection of informed consent or in the data analysis. There are no additional conflict of interests to disclose.

ORCID

Jason W. Morphew https://orcid.org/0000-0001-5971-214X

REFERENCES

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C-L., C. (1991). Effects of frequent classroom testing. *The Journal of Educational Research*, *85*, 89-99. https://doi.org/10.1080/00220671.1991.10702818

Bell, B., & Cowie, B. (2000). The characteristics of formative assessment in science education. *Science Education*, *85*, 536–553. https://doi.org/10.1002/sce.1022

Bjork, E. L., & Storm, B. C. (2011). Retrieval experience as a modifier of future encoding: Another test effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 1113–1124. https://doi.org/10.1037/a0023549

Blasiman, R. N., Dunlosky, J., & Rawson, K. A. (2017). The what, how much, and when of study strategies: Comparing intended versus actual study behavior. *Memory*, *25*, 781–792. https://doi.org/10.1080/09658211.2016.1221974

Brown, P. C., Roediger, H. L. III, & McDaniel, M. A. (2014). *Make it stick: The science of successful learning*. Cambridge, MA: Harvard University Press. https://doi.org/10.4159/9780674419377

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

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

Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H. K., & Pashler, H. (2012). Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. *Educational Psychology Review*, *24*, 369–378. https://doi.org/10.1007/s10648-012-9205-z

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science*, *19*, 1095–1102. https://doi.org/10.1111/j.1467-9280.2008.02209.x

Chan, J. C. K. (2010). Long-term effects of testing on the recall of nontested materials. *Memory*, *18*, 49–57. https://doi.org/10.1080/09658210903405737

Chan, J. C. K., Manley, K. D., Davis, S. D., & Szpunar, K. K. (2018). Testing potentiates new learning across a retention interval and a lag: A strategy change perspective. *Journal of Memory and Language*, *102*, 83–96. https://doi.org/10.1016/j.jml.2018.05.007

Chen, X. (2013). STEM attrition: College students' paths into and out of STEM fields. Statistical Analysis Report. NCES 2014-001. *National Center for Education Statistics*.

Cho, K. W., & Powers, A. (2019). Testing enhances both memorization and conceptual learning of categorical materials. *Journal of Applied Research in Memory and Cognition*, *8*, 166–177. https://doi.org/10.1016/j.jarmac.2019.01.003

Clark, I. (2012). Formative assessment: Assessment is for self-regulated learning. *Educational Psychology Review*, *24*, 205–249. https://doi.org/10.1007/s10648-011-9191-6

Cromley, J. G., Perez, T., & Kaplan, A. (2016). Undergraduate STEM achievement and retention: Cognitive, motivational, and institutional factors and solutions. *Policy Insights from the Behavioral and Brain Sciences*, *3*, 4–11. https://doi.org/10.1177/2372732215622648

Dai, T., & Cromley, J. G. (2014). Changes in implicit theories of ability in biology and dropout from STEM majors: A latent growth curve approach. *Contemporary Educational Psychology*, *39*, 233–247. https://doi.org/10.1016/j.cedpsych.2014.06.003

Darley, C. F., & Murdock, B. B. (1971). Effects of prior free recall testing on final recall and recognition. *Journal of Experimental Psychology*, *91*, 66–73. https://doi.org/10.1037/h0031836

Deck, D. W. (1998). *The effects of frequency of testing on college students in a principles of marketing course*. (Doctoral dissertation). Retrieved from http://theses.lib.vt.edu/theses/available/etd-110298-195932/unrestricted/dis.pdf.

Diegelman-Parente, A. (2011). The use of mastery learning with competency-based grading in an organic chemistry course. *Journal of College Science Teaching*, *40*, 50–58.

Dineen, P., Taylor, J., & Stephens, L. J. (1989). The effect of testing frequency upon the achievement of students in high school mathematics courses. *School Science and Mathematics*, *89*, 197–200. https://doi.org/10.1111/j.1949-8594.1989.tb11910.x

Downs, S. D. (2015). Testing in the college classroom: Do testing and feedback influence grades throughout an entire semester? *Scholarship of Teaching and Learning in Psychology*, *1*, 172–181. https://doi.org/10.1037/st10000025

Fernandez, J., & Jamet, E. (2017). Extending the testing effect to self-regulated learning. *Metacognition and Learning*, *12*, 131–156. https://doi.org/10.1007/s11409-016-9163-9

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, *31*, 799–815. https://doi.org/10.1080/0266476042000214501

Finley, J. R., & Benjamin, A. S. (2012). Adaptive and qualitative changes in encoding strategy with experience: Evidence from the test-expectancy paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 632–652. https://doi.org/10.1037/a0026215

Foss, D. J., & Pirozzolo, J. W. (2017). Four semesters investigating frequency of testing, the testing effect, and transfer of training. *Journal of Educational Psychology*, *109*, 1067–1083. https://doi.org/10.1037/edu0000197

Freeman, S., Eddy, S. L., McDnough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 8410–8415. https://doi.org/10.1073/pnas.1319030111

Hanham, J., Leahy, W., & Sweller, J. (2017). Cognitive load theory, element interactivity, and the testing and reverse testing effects. *Applied Cognitive Psychology*, *31*, 265–280. https://doi.org/10.1002/acp.3324

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin and Review*, *19*, 126–134. https://doi.org/10.3758/s13423-011-0181-y

Henderson, C., Beach, A., & Finkelstein, N. (2011). Facilitating change in undergraduate STEM Instructional practices: An analytic review of the literature. *Journal of Research in Science Teaching*, *48*, 952–984. https://doi.org/10.1002/tea.20439

Hennig, S., Staatz, C. E., Bond, J. A., Leung, D., & Singleton, J. (2019). Quizzing for success: Evaluation of the impact of feedback quizzes on the experiences and academic performance of undergraduate students in two clinical pharmacokinetics courses. *Currents in Pharmacy Teaching and Learning*, *11*, 742–749. https://doi.org/10.1016/j.cptl.2019.03.014

Johnson, B. C., & Kiviniemi, M. T. (2009). The effect of online chapter quizzes on exam performance in an undergraduate social psychology course. *Teaching of Psychology*, *36*, 33–37. https://doi.org/10.1080/00986280802528972

Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology*, *101*, 621–629. https://doi.org/10.1037/a0015183

Juhler, S. M., Rech, J. F., From, S. G., & Brogan, M. M. (1998). The effect of optional retesting on college students' achievement in an individualized algebra course. *The Journal of Experimental Education*, *66*, 125–137. https://doi.org/10.1080/00220979809601399

Kang, S. H. K., Gollan, T. H., & Pashler, H. (2013). Don't just repeat after me: Retrieval practice is better than imitation for foreign vocabulary learning. *Psychonomic Bulletin & Review*, *20*, 1259–1265. https://doi.org/10.3758/s13423-013-0450-z

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. III (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*, 528–558. https://doi.org/10.1080/09541440601056620

Karpicke, J. D., & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educational Psychology Review*, *27*, 317–326. https://doi.org/10.1007/s10648-015-9309-3

Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student learning: Do students practice retrieval when they study on their own? *Memory*, *17*, 471–479. https://doi.org/10.1080/09658210802647009

Karpicke, J. D., & Roediger, H. L. III (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 704–719. https://doi.org/10.1037/0278-7393.33.4.704

King, B. (2015). Changing college majors: Does it happen more in STEM and do grades matter? *Journal of College Science Teaching*, *44*, 44–51. https://doi.org/10.2505/4/jcst15_044_03_44

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

Kulik, C. L. C., & Kulik, J. A. (1987). Mastery testing and student learning: A meta-analysis. *Journal of Educational Technology Systems*, *15*(3), 325–345. https://doi.org/10.2190/FG7X-7Q9V-JX8M-RDJP

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

Leahy, W., Hanham, J., & Sweller, J. (2015). High element interactivity information during problem solving may lead to failure to obtain the testing effect. *Educational Psychology Review*, *27*, 291–304. https://doi.org/10.1007/s10648-015-9296-4

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science*, *23*, 1337–1344. https://doi.org/10.1177/0956797612443370

Little, J. L., Storm, B. C., & Bjork, E. L. (2011). The costs and benefits of testing text materials. *Memory*, *19*, 346–359. https://doi.org/10.1080/09658211.2011.569725

McDaniel, M. A., Agarwal, P. K., Huesler, B. J., McDermott, K. B., & Roediger, H. L. I. I. I. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, *103*, 399–414. https://doi.org/10.1037/a0021782

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

McDaniel, M. A., Roediger, H. L. I. I. I., & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review*, *14*, 200–206. https://doi.org/10.3758/BF03194052

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*, 360–372. https://doi.org/10.1002/acp.2914

McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger, H. L. I. I. I., & 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*, 3–21. https://doi.org/10.1037/xap0000004

Nelson, C. E. (1996). Student diversity requires different approaches to college teaching, even in math and science. *American Behavioral Scientist*, *40*, 165–175. https://doi.org/10.1177/0002764296040002007

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin*, *144*, 710–756. https://doi.org/10.1037/bul0000151

Paternoster, R., Brame, R., Mazerolle, P., & Piquero, A. (1998). Using the correct statistical test for the equality of regression coefficients. *Criminology*, *36*, 859–866. https://doi.org/10.1111/j.1745-9125.1998.tb01268.x

Peterson, D. J., & Wissman, K. T. (2018). The testing effect and analogical problem-solving. *Memory*, *26*, 1460–1466. https://doi.org/10.1080/09658211.2018.1491603

Pikunas, J., & Mazzota, D. (1965). The effects of weekly testing in the teaching of science. *Science Education*, *49*, 373–376. https://doi.org/10.1002/sce.3730490415

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, *330*, 335. https://doi.org/10.1126/science.1191465

Rask, K. (2010). Attrition in STEM fields at a liberal arts college: The importance of grades and pre-collegiate preferences. *Economics of Education Review*, *29*, 892–900. https://doi.org/10.1016/j.econedurev.2010.06.013

Rawson, K. A., & Dunlosky, J. (2012). When is practice testing most effective for improving the durability and efficiency of student learning? *Educational Psychology Review*, *24*, 419–435. https://doi.org/10.1007/s10648-012-9203-1

Rawson, K. A., Dunlosky, J., & Sciartelli, S. M. (2013). The power of successive relearning: Improving performance on course exams and long-term retention. *Educational Psychology Review*, *25*, 523–548. https://doi.org/10.1007/s10648-013-9240-4

Richland, L. E., Kao, L. S., & Kornell, N. (2008). Can unsuccessful tests enhance learning? In *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society* (pp. 2338–2343). Austin, TX: Cognitive Science Society.

Roediger, H. L. III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*, 20–27. https://doi.org/10.1016/j.tics.2010.09.003

Roediger, H. L. III, & Karpicke, J. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*, 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Rohrer, D., Taylor, K., Pashler, H., Cepeda, N. J., & Wixted, J. T. (2005). The effect of overlearning on long-term retention. *Applied Cognitive Psychology*, *19*, 361–374. https://doi.org/10.1002/acp.1083

Roszkowski, M. J., & Speat, S. (2016). Retaking the SAT may boost scores but this doesn't hurt validity. *Journal of the National College Testing Association*, *2*, 1–16.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*, 207–217. https://doi.org/10.1111/j.1467-9280.1992.tb00029.x

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods*, *11*, 54–71. https://doi.org/10.1037/1082-989X.11.1.54

Sobek, D. K., Cundy, V. A., & Briggeman, V. L. (2004). Assessing the given–find–solution method in an undergraduate thermodynamics course. *International Journal of Mechanical Engineering Education*, *32*, 183–196. https://doi.org/10.7227/IJMEE.32.3.1

Soderstrom, N. C., & Bjork, R. A. (2014). Testing facilitates the regulation of subsequent study time. *Journal of Memory and Language*, *73*, 99–115. https://doi.org/10.1016/j.jml.2014.03.003

Son, L. K., & Simon, D. A. (2012). Distributed learning: Data, metacognition, and educational implications. *Educational Psychology Review*, *24*(3), 379–399. https://doi.org/10.1007/s10648-012-9206-y

Stephens, L. J. (1977). The effect of the class evaluation method on learning in certain mathematics courses. *International Journal of Mathematical Education in Science and Technology*, *8*, 477–479. https://doi.org/10.1080/0020739770080414

Thomas, R. C., Weywadt, C. R., Anderson, J. L., Martinez-Papponi, B., & McDaniel, M. A. (2018). Testing encourages transfer between factual and application questions in an online learning environment. *Journal of Applied Research in Memory and Cognition*, *7*, 252–260. https://doi.org/10.1016/j.jarmac.2018.03.007

Townsend, N. R., & Wheatley, G. H. (1975). Analysis of frequency of tests and varying feedback delays in college mathematics achievement. *College Student Journal*, *9*, 32–36.

Van Gog, T., & Kester, L. (2012). A test of the testing effect: Acquiring problem-solving skills from worked examples. *Cognitive Science*, *36*, 1532–1541. https://doi.org/10.1111/cogs.12002

Van Gog, T., Kester, L., Dirkx, K., Hoogerheide, V., Boerboom, J., & Verkoeijen, P. P. J. L. (2015). Testing after worked example study does not enhance delayed problem-solving performance compared to restudy. *Educational Psychology Review*, *27*, 265–289. https://doi.org/10.1007/s10648-015-9297-3

Van Gog, T., Kester, L., & Paas, F. L. (2011). Effects of worked examples, example–problem, and problem–example pairs on novices' learning. *Contemporary Educational Psychology*, *36*, 212–218. https://doi.org/10.1016/j.cedpsych.2010.10.004

Van Gog, T., & Sweller, J. (2015). Not new, but nearly forgotten: The testing effect decreases or even disappears as the complexity of learning materials increases. *Educational Psychology Review*, *27*, 247–264. https://doi.org/10.1007/s10648-015-9310-x

Vaughn, K. E., & Rawson, K. A. (2011). Diagnosing criterion-level effects on memory: What aspects of memory are enhanced by repeated retrieval? *Psychological Science*, *22*, 1127–1131. https://doi.org/10.1177/0956797611417724

Ward, E. F. (1984). Statistics mastery: A novel approach. *Teaching of Psychology*, *11*, 223–225. https://doi.org/10.1177/009862838401100409

William, B. D., House, W. J., & Boyd, T. L. (1984). A test–retest policy for introductory psychology courses. *Teaching of Psychology*, *11*, 182–184.

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2011). The interim test effect: Testing prior material can facilitate the learning of new material. *Psychonomic Bulletin & Review*, *18*, 1140–1147. https://doi.org/10.3758/s13423-011-0140-7

Wooldrige, 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*, 214–221. https://doi.org/10.1016/j.jarmac.2014.07.001

Zilles, C., West, M., Mussulman, D., & Bretl, T. (2018). *Making testing less trying: Lessons learned from operating a computer-based testing facility*. Proceedings of the 2018 Frontiers in Education Conference (FIE2018).

How to cite this article: Morphew JW, Silva M, Herman G, West M. Frequent mastery testing with second-chance exams leads to enhanced student learning in undergraduate engineering. *Appl Cognit Psychol*. 2020;34:168–181. https://doi.org/10.1002/acp.3605