Computers & Education 237 (2025) 105387

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

# The promise of mastery-based testing for promoting student engagement, self-regulated learning, and performance in gateway STEM courses

Michael W. Asher[a,*,1], Joshua D. Hartman[b,**,1], Mark Blaser[c], Jack F. Eichler[b], Paulo F. Carvalho[a]

[a] *Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA, USA*
[b] *Department of Chemistry, University of California, Riverside, Riverside, CA, USA*
[c] *Chemistry Department, Shasta College, Redding, CA, USA*

ARTICLE INFO

*Keywords:*
Self-regulated learning
Mastery learning
Computer-mediated learning
Educational data mining

ABSTRACT

Decades of research show that tests, beyond assessing student knowledge, are powerful tools for promoting learning. However, high-stakes tests can also cause stress and disengagement. To utilize tests to encourage and motivate students, we implemented a mastery-based testing system in a large-enrollment general chemistry course ($N = 234$). This system allowed students to take three versions of each unit test, studying digital course resources in between to increase their mastery of the content. Students chose to take advantage of the mastery testing system when they struggled with unit tests, averaging six total repeated attempts. This level of repeated testing was associated with a 60 % increase in students' use of online study resources over the duration of the course, and a five-point overall increase in final exam scores (11 points for first-generation college students). Critically, student engagement with digital learning materials mediated these performance gains, suggesting that the benefits of mastery-based testing systems were not only due to students responding to the tests themselves. Instead, the findings suggest that mastery-based testing systems can enhance performance in introductory STEM courses by providing motivation and structure to support students' self-regulated learning, helping them invest more time in effective, distributed study strategies.

## 1. Introduction

Although tests are often seen as tools for assessing what students know or evaluating their performance, decades of research show that tests can also be helpful tools for instruction. When students are tested on previously learned information, the process of retrieving the information from memory can strengthen long-term retention more effectively than re-reading (Carpenter et al., 2022; Roediger & Karpicke, 2006; Rowland, 2014). "Pretests" given before readings or lectures have also been shown to enhance learning by triggering students' interest and orienting their attention to relevant information (Pan & Carpenter, 2023). Moreover, tests provide critical

* Corresponding author.
** Corresponding author.
*E-mail addresses:* masher@andrew.cmu.edu (M.W. Asher), joshua.hartman@ucr.edu (J.D. Hartman).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.compedu.2025.105387
Received 29 January 2025; Received in revised form 7 May 2025; Accepted 19 June 2025
Available online 19 June 2025

metacognitive benefits. Passive learning approaches can create a false sense of understanding; when students listen to a lecture or read a text without having their understanding challenged, they are likely to experience a sense of fluency, comprehension, and therefore overconfidence (Bjork & Bjork, 2011). When students engage with the material through testing, they uncover gaps in their knowledge, allowing them to better assess what they have mastered and what needs further study (Bjork et al., 2013; Kornell et al., 2009; Pan & Carpenter, 2023).

Evidence for the importance of testing is found in laboratory experiments in the cognitive science literature on topics like the testing effect, pre-questioning, and successive re-learning (Carpenter & Toftness, 2017; Carrier & Pashler, 1992; Rawson et al., 2018). For example, a meta-analysis of laboratory studies found that students who practiced retrieving information performed, on average, about half a standard deviation better on later recall tests than students who simply restudied the material for the same amount of time (Rowland, 2014). A meta-analysis of classroom experiments found a slightly smaller, but still meaningful, benefit of practice testing—about three-tenths of a standard deviation (Yang et al., 2021). Further evidence comes from analyses of large-scale online courses, where students' study behaviors can be precisely tracked. In these settings, the associations between practice testing and learning outcomes have been estimated to be two to six times stronger than those between passive instructional activities (e.g., reading or watching videos) and learning (Carvalho et al., 2022; Koedinger et al., 2015, pp. 111–120).

In addition to providing students with the critical practice opportunities they need to learn, graded tests can motivate students to engage outside of class. Although it is important for educators to intrinsically motivate their students (e.g, by developing interesting lessons and demonstrating the value of course content), extrinsic rewards like grades also remain an important tool (Cerasoli et al., 2014). Grades can be particularly effective at encouraging deeper engagement with course content for students who struggle or are initially unmotivated (Hidi & Harackiewicz, 2000).

However, graded tests also have significant potential downsides that must be considered and addressed. Graded tests cause anxiety for many students (Von Der Embse et al., 2018), which can lead them to underperform and disengage (Beilock et al., 2004; Chamberlin et al., 2023). This risk can be greatest for students who belong to negatively stereotyped groups and fear that failure will confirm the stereotypes (see the literature on "stereotype threat," e.g., Steele, 1997). In addition, graded tests can create a classroom environment in which students become preoccupied with avoiding failure in front of others (Pulfrey et al., 2011). These "performance-avoidance goals" can lead struggling students to disengage from a course to save face and avoid stress (Elliot & Harackiewicz, 1996), dodging the critical practice opportunities they need to learn and succeed.

One way to mitigate these risks while maintaining the benefits of retrieval practice is through low-stakes assessments, such as ungraded quizzes (e.g., McDaniel et al., 2011). These approaches allow students to test and reinforce their knowledge without fear of a substantial penalty, promoting learning while minimizing anxiety. However, because low-stakes assessments lack consequences, they have also been associated with students putting forth less effort (Penk & Richter, 2017).

We predict that mastery-based testing systems, which allow students to improve their test scores with repeated attempts, will be a powerful way to build upon the benefits of low-stakes assessment while also serving as a productive motivator, shifting the focus from performance on a single test to long-term learning and improvement. In the present research, we examine if a mastery-based testing system can motivate effective self-regulated studying in a general chemistry course (Research Question 1) and thereby improve learning and performance (Research Question 2).

### *1.1. Theoretical framework*

Mastery-based testing is closely related to the broader idea of mastery-based grading. Mastery-based grading systems share the following features (Fernandez, 2021).

1. Assignments are based on clearly articulated objectives (e.g., a specific skill or piece of the curriculum).
2. Students are graded based on their mastery of the objectives.
3. Students are given multiple opportunities to repeat assignments and thereby demonstrate mastery.

Whereas mastery-based grading can be used with any type of assignment (see "specification grading" systems in chemistry labs, Howitz et al., 2021; Ring, 2017), we define mastery-based testing as a grading system in which *tests* are centered around learning objectives, students receive grades and feedback on their initial performance, and students have the opportunity to retake similar tests on the same objectives to demonstrate further mastery and improve their grades.

By administering mastery-based tests, instructors might capitalize on the cognitive and metacognitive benefits of repeated practice testing in a way that promotes sustained engagement and reduces motivation-related risks. By providing repeated, lower-stakes opportunities for students to gauge their knowledge, address their weaknesses, and retake tests to demonstrate improved mastery, mastery-based testing systems may offer a powerful incentive for students to engage with practice materials over time without the fear of failure.

Critically, these systems should allow struggling students the chance to complete additional practice attempts needed to catch up. Furthermore, mastery-based testing may incentivize distributed practice as students return to the same content when they repeatedly prepare for and take assessments over time. For example, if instructors require a one-week interval between testing attempts, each attempt provides a form of spaced retrieval. These repeated tests should also encourage students to engage in further self-directed, spaced practice—such as working through practice problems from earlier textbook units—in preparation for subsequent attempts. As such, mastery-based testing can help students avoid "cramming" sessions, which are ineffective for promoting long-term memory (see research on spacing and interleaving, e.g., Carpenter et al., 2022; Firth et al., 2021).

Fig. 1 illustrates how mastery-based testing systems might enhance learning over time. In a course with traditional testing, tests serve as summative assessments at the end of a topic or unit. Although they motivate students to study in preparation, students often have little incentive to review mistakes or revisit content afterward. Engagement typically ends after completing the test, the third box in Fig. 1. In contrast, the additional opportunities provided by mastery-based testing should encourage students to review and correct their mistakes as they prepare for subsequent attempts. These extra opportunities allow students to channel their performance goals into an adaptive focus on studying and improvement (Senko et al., 2011).

Thus, there are multiple pathways by which mastery-based testing might improve learning. First, by enabling repeated, spaced practice through testing itself, mastery-based tests could directly facilitate encoding, memory, and generalization. Second, by providing a motivating and organizing framework, mastery-based systems have the potential to indirectly promote spaced retrieval by facilitating effective self-regulated learning. This structure could guide students' metacognition and drive them to plan targeted, distributed study sessions over time. Fig. 2 summarizes the process of self regulated learning, which is often characterized as a cycle of reflection, planning, and performance (Panadero, 2017), and the figure also provides an overview of the hypothesized benefits of mastery-based testing.

Quantitative evaluations of mastery-based testing systems are rare; however, initial evidence is promising. In a quasi-experimental evaluation of a "second-chance" testing system in an undergraduate engineering course, Morphew and colleagues (2020) integrated seven computer-based tests into the curriculum, provided feedback after each test, and allowed students to retake a similar test the following week. Compared to students who took the same instructor's course the semester before the mastery-based testing system was implemented, students' final exam performance improved by seven percentage points (*Cohen's d* = .50), suggesting that this grading structure can facilitate longer-term improvements in learning.

In the present study, we examine the longitudinal links between the use of mastery-based testing and student engagement, learning, and performance. To do so, we implemented a mastery testing system during the second quarter of the three-part general chemistry sequence at the University of California, Riverside, a public university and federally designated Hispanic Serving Institution.

### 1.2. Hypotheses

The following three hypotheses guided our research.

**H1.** The mastery-based testing system will be utilized primarily by struggling students.

**H2.** Use of the mastery-based testing system will correlate with students studying the contents of previous units outside of class (as they prepared to repeat tests throughout the semester).

**H3.** Engagement with the mastery-based testing system (and associated studying) will be associated with learning and achievement, both on the repeated tests themselves and on the course's final exam.

## 2. Materials and methods

This study was approved by the Institutional Review Board at the university of California, Riverside. Data were collected in *Chemistry 1B,* the second course in a three-part general chemistry sequence. To enroll in this course, students were required to have completed *Chemistry 1A*—the first course in the sequence, offered in the first quarter of the academic year—with a grade of C- or higher. *Chemistry 1A* was a distinct course with no mastery-based testing system. All students were enrolled in the same section of the course, which was taught by one instructor with a combination of face-to-face lectures, an online textbook with interactive practice problems to provide additional instruction (see Lovett et al., 2008), and proctored, digital assessments. In addition, students who took the general chemistry course also were required to enroll in a laboratory section. Although this laboratory section was offered as a

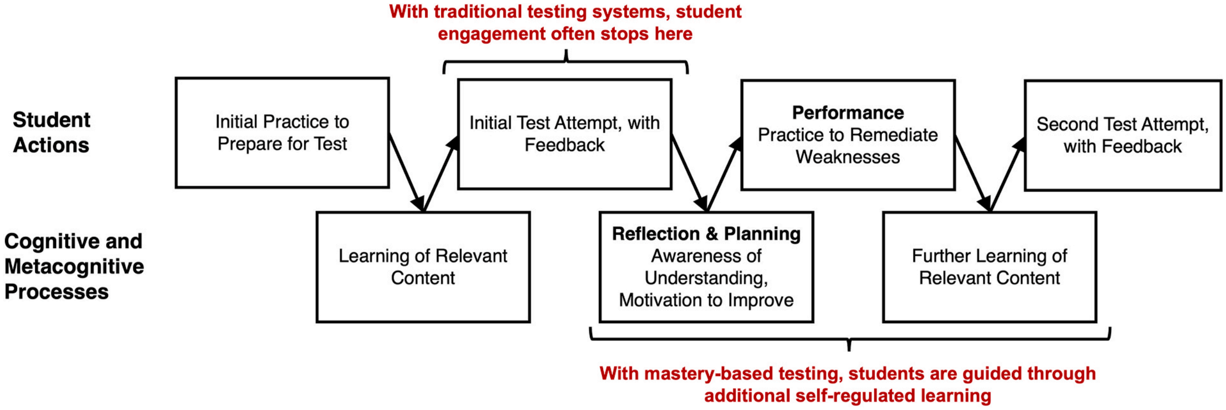
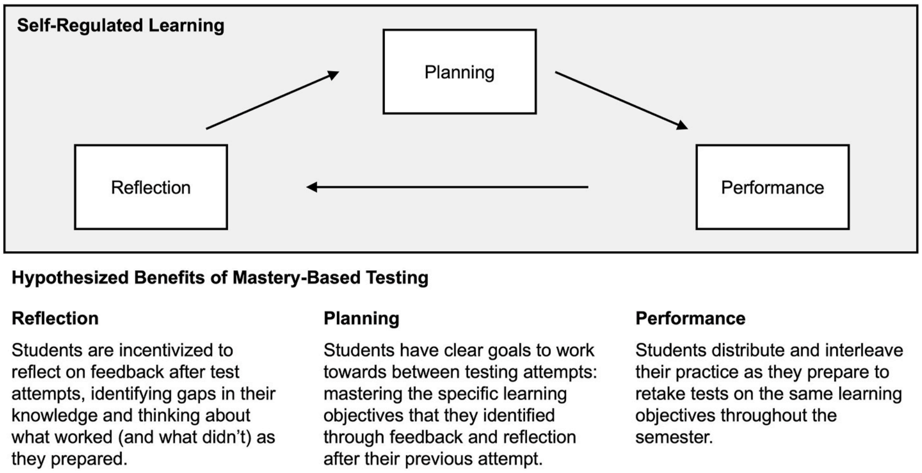

With traditional testing systems, student engagement often stops here
Student Actions
Initial Practice to Prepare for Test
Initial Test Attempt, with Feedback
Performance
Practice to Remediate Weaknesses
Second Test Attempt, with Feedback
Cognitive and Metacognitive Processes
Learning of Relevant Content
Reflection & Planning
Awareness of Understanding, Motivation to Improve
Further Learning of Relevant Content
With mastery-based testing, students are guided through additional self-regulated learning

**Fig. 1.** Conceptual model of mastery-based testing.



**Hypothesized Benefits of Mastery-Based Testing**

**Reflection**

Students are incentivized to reflect on feedback after test attempts, identifying gaps in their knowledge and thinking about what worked (and what didn't) as they prepared.

**Planning**

Students have clear goals to work towards between testing attempts: mastering the specific learning objectives that they identified through feedback and reflection after their previous attempt.

**Performance**

Students distribute and interleave their practice as they prepare to retake tests on the same learning objectives throughout the semester.

**Fig. 2.** Hypothesized benefits of mastery-based testing for self-regulated learning.

separate course, it was closely aligned with the lecture and exams in terms of content and pacing.

### *2.1. Participants*

In total, 244 students were enrolled in the course, 234 of whom used the online courseware and completed the course and therefore comprise the sample for the current study. Of these students, 82 (35 %) were first-generation college students. There were 122 (52 %) Asian or Asian American students, 66 (28 %) Hispanic or Latino students, 22 White students (9 %), 9 multiracial students (4 %), 4 Black students (2 %), and 13 students (6 %) who did not report their race/ethnicity. Information on age and gender was not available due to IRB restrictions.

### *2.2. The mastery-based testing system*

The general chemistry course was divided into six units, each covering a distinct topic (e.g., gasses, thermochemistry). Each unit lasted for 1–2 weeks. To implement a mastery-based testing system, we introduced a 10-question assessment at the end of each unit to evaluate students' understanding of the unit's learning objectives. These six tests accounted for 60 % of the students' overall course grades, 10 % each. We created a question bank with up to 500 questions for each unit, each tagged with a specific learning objective. From this bank, we sampled questions to generate three parallel versions of each unit's test, with each version containing problems for the same learning objectives. Next, the course's instructor reviewed the tests to ensure that they covered the same topics and were approximately the same difficulty. Test items were designed to assess students' knowledge of course concepts and their ability to solve problems. Sample items are presented in the Supplementary Materials. Students were given 25 min to complete each test. Tests were administered in person so they could be proctored, and they were delivered via the *Canvas* learning management system to facilitate automated grading and feedback.

Immediately after students completed a unit's test, they received a score along with feedback on their responses, including whether each answer was correct or incorrect and the correct answers for any mistakes (see Pashler et al., 2005). Students were then allowed to repeat the test up to two more times, keeping only their highest score. As such, students could take as many as 18 total tests during the semester. They were required to complete each test once (six first attempts) and could optionally complete two more attempts for each test (12 repeated attempts). Students were required to wait a minimum of one week between attempts for each test to give them time to study. In addition to these six tests that were graded with a mastery-based policy, the course also had traditional midterm and final exams, each worth 10 % of the final grade. However, higher scores on these exams could also replace lower grades on unit mastery tests with corresponding content.

To study for course exams (on both first and repeated attempts) students were encouraged to use General Chemistry courseware developed by the Open Learning Initiative (OLI) at Carnegie Mellon University. Like a traditional textbook, this online resource provided readings and practice exercises for each of the course's six units. Unareflike a traditional textbook, however, the OLI platform offered an interactive learning environment with real-time support and immediate feedback (Bier et al., 2019). While some research suggests that delayed feedback can be beneficial when working memory is heavily taxed (Schooler & Anderson, 1990, pp. 702–708), a broader body of evidence supports the effectiveness of immediate feedback in educational contexts, including technology-based learning environments (Dabbagh et al., 2019; Kulik & Kulik, 1988). In addition, a recent large-scale study comparing immediate and delayed feedback found no difference in real-world learning outcomes (Fyfe et al., 2021), suggesting that immediate feedback is at least as effective in most cases.

OLI's content included textbook passages, checkpoint quizzes (students received 15 % of their overall grade for completing these checkpoints for homework), review questions, and scaffolded feedback designed to support conceptual understanding and problem-solving. These checkpoints and practice questions were aligned with the same learning objectives as the mastery-based unit tests and course exams. However, there were differences in format: the assessments (including unit and final exams) consisted primarily of multiple choice and numerical response items, while the OLI courseware incorporated a wider variety of problem types—including matching, fill-in-the-blank, and multiple select questions—accompanied by hints and elaborated feedback. This variety of formats allowed students to engage with the material in different ways and receive tailored support as they prepared for assessments. Sample OLI questions are shared in the Supplementary Materials. In addition to delivering real-time instructional feedback, the OLI platform also logged detailed records of student activity, enabling precise measurement of out-of-class study behavior.

### *2.3. Measures*

***Repeated Test Attempts.*** To capture use of the mastery-based testing system, we counted the number of repeated unit tests that each student attempted ($M = 5.74$, $SD = 2.44$). Each test could be attempted three times by each student, and as such students could complete a maximum of 12 repeat attempts.

***Chemistry Problems Completed before First Test Attempts.*** The chemistry course used a combination of the “General Chemistry I” and “General Chemistry II” chemistry courseware produced by OLI. To measure students’ outside-of-class study behavior as they prepared for their first test attempts, we tracked the number of OLI chemistry problems that each student completed in each unit before attempting the corresponding unit test for the first time ($M = 583.5$, $SD = 380.3$).

***Chemistry Problems Completed after First Test Attempts.*** To measure the extent that students revisited practice problems from previous units’ later in the course (e.g., while preparing to retake mastery-based tests) we also tracked the number of OLI problems completed by each student in each unit after attempting the corresponding unit test for the first time ($M = 996.2$, $SD = 467.1$). These problems could represent a student preparing for a repeated mastery-based test, but they could also have taken place after all mastery-based testing attempts and show a student studying for a midterm or final exam.

***Chemistry Problems Completed in Between Mastery Test Attempts.*** Finally, to more closely measure the impact of mastery-based testing on engagement with course content, we counted the number of total OLI problems that students completed in corresponding units between mastery test attempts ($M = 574.5$, $SD = 355.3$). For example, for this measure we only counted the problems that were attempted by each student in Unit 1 after their first mastery-based test for Unit 1, but before any second and third attempts that they took of this exam.

***Mastery-Based Test Grades.*** We tracked students’ performance across all attempts on the six mastery-based unit tests (M = 58 %, SD = 23 %). Test-retest reliability for these assessments was relatively low (ICC = .34), suggesting that students were somewhat consistent in their performance throughout the year, but that the mastery-based tests also allowed for substantial within-student change over time.

***Final Exam Grade.*** At the end of the semester, students took a cumulative final exam for the class that covered content from all six units of the course. Whereas the unit tests were 10 questions each and students were expected to finish in 25 min, the final exam had 32 questions, and students were given 2 h to finish. The final exam had good internal consistency, $\alpha = .82$, and students answered 71.7 % of exam questions correctly ($SD = 16.4$ %). The final exam grades from this course were also analyzed as part of a multi-course study about the implementation of different versions of mastery testing in general chemistry (Hartman & Eichler, 2024).

## 3. Results

All analyses were conducted using R version 4.4.1 (R Core Team, 2024). The detailed output (e.g., coefficients, standard errors, and inferential statistics) for all models is reported in the Supplementary Materials, along with descriptive statistics. Data and code to reproduce all analyses and figures are shared at https://osf.io/6pw9k.

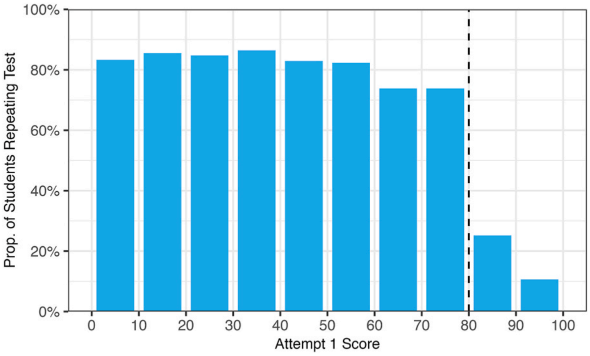


**Fig. 3.** First attempt scores vs. the proportion of students repeating each unit test.

### 3.1. *Struggling students were more likely to repeat tests*

The relationship between (a) the first-attempt scores for all students on all mastery tests, and (b) whether students chose to repeat each test is displayed in Fig. 3. As predicted, students who struggled on their first attempt were much more likely to repeat the test: a one standard-deviation decrease in first attempt scores was associated with a more than 3x increase in the odds of a student repeating the test at least once, $p < .001$. Students were likely to repeat unit tests at least once if they attained lower than an 80 % on their first attempt (with 81 % of students doing so), but only 21 % of students who attained a score greater than 80 % chose to retake that test. A horizontal line has been added to Fig. 3 to emphasize this criterion that students appeared to impose on themselves.

### 3.2. *Use of the testing system was associated with students returning to practice course resources from previous units*

On average, students chose to repeat approximately six tests over the duration of the course. Fig. 4 shows the relationship between the number of repeated tests that each student took and the number of problems that they completed in relevant units of the online courseware, both before (4A) and after (4B) attempting each test for the first time. Although usage of the mastery testing system was unrelated to the number of problems that students completed before their first attempts, $t(232) = .506$, $p = .613$, each additional repeated test was associated with a student circling back to a previous unit and completing an additional 95.4 practice problems after their first test attempt, $t(232) = 8.75$, $p < .001$, an approximately 10 % increase in total studying per repeated attempt. This pattern of results is consistent with the mastery testing system encouraging students to return to course materials to prepare themselves for subsequent mastery attempts, increasing the total studying students completed.

### 3.3. *Studying between mastery attempts was associated with improved performance on repeated tests*

What happened when students decided to retake a unit test and revisit that unit's contents? Was the additional practice associated with learning? As Fig. 5 shows, studying relevant content between test attempts was associated with improved performance; each additional 100 practice problems completed predicted a 5.1 point improvement in test scores between attempts one and two, $F(1, 965) = 45.77$, $p = .001$ (Fig. 5A), and a 4.9 point improvement between attempts two and three, $F(1, 321) = 5.01$, $p = .025$ (5B). Notably, this analysis suggested that students were unlikely to show improvement in test scores unless they studied between attempts.

### 3.4. *Use of the mastery testing system was associated with final exam performance, with effects mediated via study behavior*

Next, we examined (a) how usage of the mastery testing system was related to end-of-semester performance on the final exam, and (b) whether studying between mastery-test attempts would mediate any effects of using the system on end-of-semester performance. Because struggling students were more likely to utilize the mastery testing system, we controlled for students' average baseline performance on unit tests in these analyses to examine the marginal impact of utilizing the mastery testing system and studying between attempts, holding initial performance constant.

Overall, choosing to repeat a single unit test was associated with a .87-point increase in a student's final exam grade, controlling for their average initial performance, $t(231) = 2.20$, $p = .029$. Given that students averaged six total repeated attempts, use of the mastery-based testing system was associated with an average of a five-point increase in final exam scores. Completing additional practice problems between test attempts was also associated with end-of-semester exam grades, suggesting the association between use of the mastery testing system and final grades might be mediated by students increasing their engagement with practice problems.

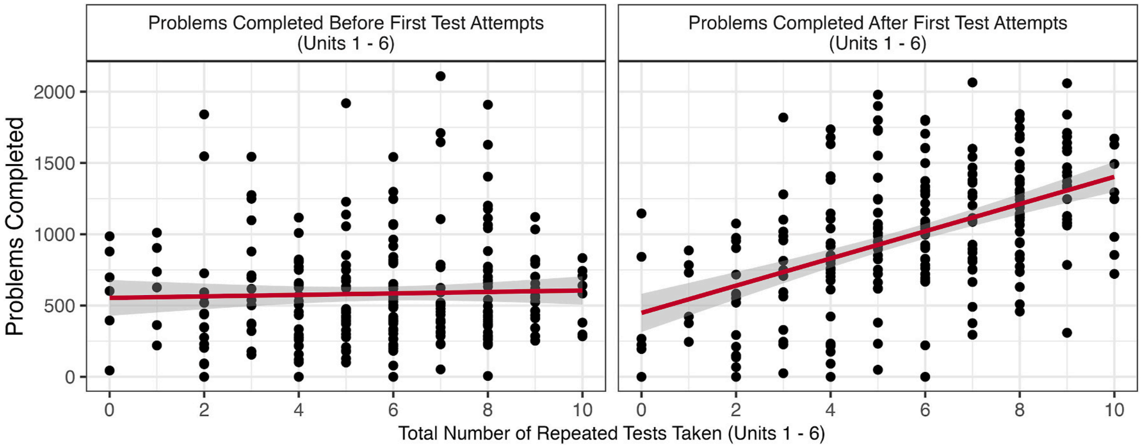


**Fig. 4.** Use of the mastery testing system vs. online chemistry problems completed in all units before (4A) and after (4B) the first corresponding mastery-based test. Plotted lines show predicted values from the regression models used to test the relationships. Error envelopes represent ± 1 standard error.

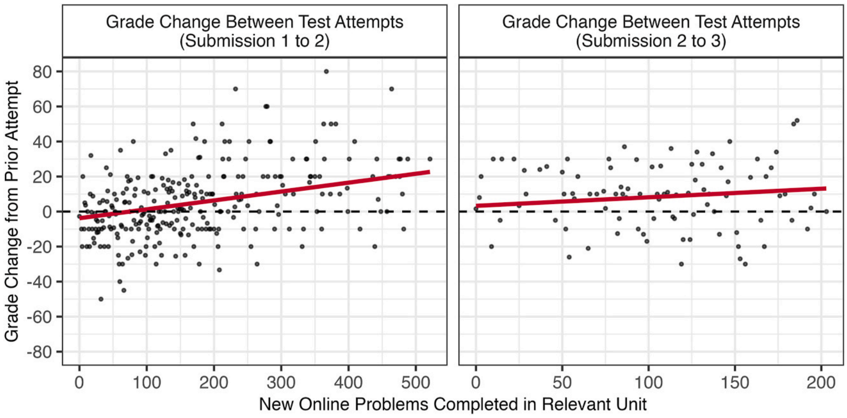


**Fig. 5.** Problems completed between repeated tests, compared with change in performance between attempts. Each point represents one unit test for each student. The plotted lines show predicted values from the linear mixed-effects models used to test the relationship for submissions one and two (5A) and submissions two and three (5B).

Specifically, each additional problem completed between mastery-test attempts predicted a .09-point increase in final exam scores, $t(231) = 8.49$, $p < .001$.

To test for mediation, we fit the path model depicted in Fig. 6, again controlling for students' average baseline performance on unit tests and using a bootstrapping procedure to compute the indirect effect (i.e., the path through the mediator, Preacher & Hayes, 2004). In this model, we found that accounting for the number of problems that students completed between repeated tests reduced the relationship between repeated tests and final exam performance by .95 units, 95 % CI = [0.16, 1.72]. This suggests that the association between usage of the mastery testing system and final exam performance was driven by students studying to prepare for the repeated tests.

### *3.5. The association between use of the mastery testing system and final exam performance was marginally stronger for first-generation college students*

Finally, we explored whether use of the mastery testing system was associated with more equitable outcomes in the chemistry course. On average, first-generation college students (who are the first in their family to attend a four-year college) arrive at college with less prior preparation for the specific demands of studying for college-level science courses. For instance, first-generation college students are less likely to have taken advanced-placement science courses in high school (Warburton et al., 2001), and tend to have fewer connections who they can turn to for help with college-related issues (Nichols & Islas, 2016). Accordingly, we reasoned that the mastery testing system might particularly help these students by providing a structure to support their self-regulated learning.

Fig. 7 shows the relationship between use of the mastery testing system and final exam performance for first- and continuing-generation college students. The interaction between use of the mastery testing system and first-generation status was non-significant, $t(229) = -1.80$, $p = .073$, but the results followed the predicted pattern: among students who never used the mastery testing system, the model predicted a 12.2 point difference in final exam performance between first- and continuing-generation students, with many continuing-generation students demonstrating above-average performance with no repeated tests. However, among first-generation college students, each additional repeated test was associated with a 1.8-point increase in final exam grades, $t(229) = 2.76$, $p = .006$. Thus, the model suggests an 11-point increase in final exam grades for first-generation students who repeat six tests (the average for the course), narrowing the achievement gap.

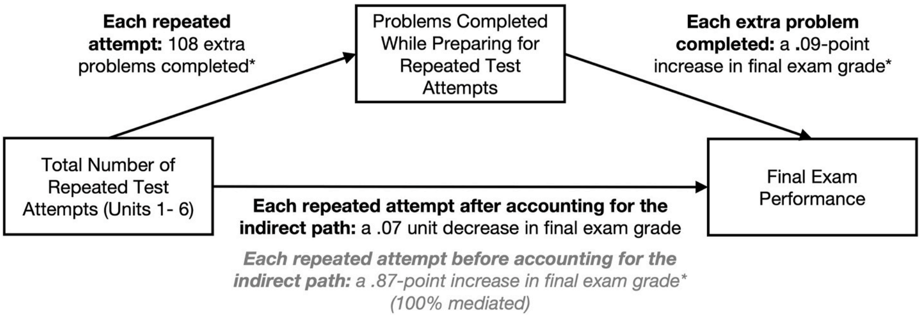

**Each repeated attempt:** 108 extra problems completed*
Problems Completed While Preparing for Repeated Test Attempts
**Each extra problem completed:** a .09-point increase in final exam grade*
Total Number of Repeated Test Attempts (Units 1- 6)
Final Exam Performance
**Each repeated attempt after accounting for the indirect path:** a .07 unit decrease in final exam grade
***Each repeated attempt before accounting for the indirect path:*** *a .87-point increase in final exam grade**
*(100% mediated)*

**Fig. 6.** Study behavior mediated the mastery testing system's association with final exam performance. Asterisks indicate a path is statistically significant.

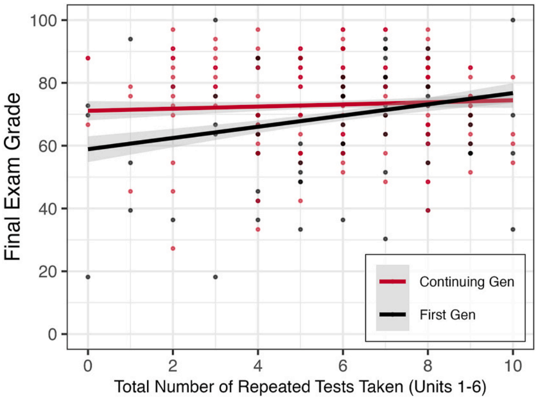


**Fig. 7.** Final exam grades vs. repeated mastery tests for first- and continuing-generation college students. Plotted lines show predicted values from the regression model used to test the relationships. Error envelopes represent ± 1 standard error.

## 4. Discussion

After implementing a mastery-based testing system in a general chemistry course–which gave students the opportunity to repeat each of the six unit tests up to three times–we found evidence that this system could effectively leverage the benefits of testing to promote learning. Although the course already included multiple components that promoted distributed practice of course content (including checkpoint quizzes, lab sessions, and repeated problem-solving exercises) mastery-based testing appeared to motivate additional spaced practice. Students took advantage of the mastery system when they struggled with unit tests, averaging six total repeated attempts. This level of repeated testing was associated with a 60 % increase in students' use of study resources over the duration of the course, and a five-point overall increase in final exam scores, with potentially larger benefits for first-generation college students. Our findings indicate that the system's positive impact on learning outcomes was mediated by increased engagement with online practice resources, reflecting self-regulated studying. Although we did not collect survey measures of students' motivation during the course, the strong link between the mastery-testing system and studying provides behavioral evidence that the system motivated students to engage with course content over time.

These results reinforce the well-documented direct benefits of testing for learning, as students demonstrated improvement through repeated practice. In addition, mediation analysis suggests that the primary benefits of mastery-based testing for learning may be indirect, driven by the system motivating and supporting self-regulated learning.

### *4.1. Direct and indirect benefits of practice testing*

Previous research on test-enhanced learning has primarily focused on the direct benefits of testing, for example for memory, metacognitive awareness, and generalization (Carpenter et al., 2022; Pan & Rickard, 2018; Roediger & Karpicke, 2006). However, our research also suggests two *indirect* benefits of testing, at least when implemented in a lower-stakes, mastery-oriented manner. First, testing can encourage students to engage more deeply with relevant material as they prepare for tests, motivated by the desire to improve their performance and grades. Students in this study were much more likely to complete practice problems for previous units when they were preparing for repeated mastery tests, creating helpful opportunities for spaced practice of previously-studied material. From a theoretical perspective this motivation-related benefit should not be overlooked, because it contrasts with decades of theory that tests are demotivating (e.g., by being controlling and anxiety-provoking, Ryan & Weinstein, 2009) and research showing that high-stakes tests can cause students to disengage from their courses (Chamberlin et al., 2023).

Second, the observed increases in outside-of-class engagement suggest that testing can provide structure to support the three-phases of self-regulated learning (Panadero, 2017), motivating students to reflect on their understanding of course content, helping them focus on specific goals for study, and providing weekly testing opportunities that hold students accountable and help them study with distributed, interleaved practice. In addition, the digital, personalized support provided by the OLI platform may have been critical for helping students make progress as they studied. Without access to effective resources, self-regulated study can be unsuccessful (Kirschner et al., 2006), and the hints and feedback from intelligent tutoring systems have been shown to promote learning for students with lower levels of prior knowledge (Koedinger et al., 2023). As such, personalized support is likely helpful to ensure the success of mastery-based testing for struggling students. Computer-based testing also played a critical role in implementing a mastery-based testing system at scale, as it allowed for automated grading and the provision of timely feedback.

### *4.2. Potential benefits and pitfalls for student motivation*

An important next step for research on mastery-based testing will be to examine the possible downstream consequences of this teaching practice on student motivation. As previously discussed, high-stakes testing is rife with potential negative consequences, causing stress, replacing intrinsic with extrinsic motivation, undermining students' beliefs about their ability and potential, and focusing students' attention on avoiding failure rather than growing and learning. These consequences can thwart students' interest in a subject and cause them to disengage (Ryan & Weinstein, 2009). The behavioral measures of engagement that we examine in this work suggest that mastery-based testing was not demotivating, at least when it came to students' choices to engage with course content, on average. However, it is critical to examine motivation in more detail over a longer time frame, supplementing behavioral data with self-report measures of constructs such as interest, self-efficacy, self-concept, achievement goals, learning strategies, intentions, and anxiety.

Mastery testing might differ from high-stakes testing in ways that provide motivational benefits: by giving students multiple opportunities to improve their grade on each assessment, teachers lower the stakes of testing in their class, potentially reducing stress and threat. Additionally, when a teacher adopts mastery testing this communicates that they believe their students' ability can grow with practice (Kroeper et al., 2022). Furthermore, if students are incentivized to study and improve through lower-stakes testing, they could discover that they can overcome academic challenges with effective practice strategies, developing a robust sense of self-efficacy (Bandura, 1997). By increasing the number of interactions students have with course materials, teachers also increase the number of opportunities students have to learn something that they find interesting or relevant. Overall, students who take part in an introductory science course with mastery-based testing might build the skills and motivational beliefs that they need to be engaged, effective self-regulated learners and persist in college science courses.

Despite these potential benefits, for many students, even mastery-based tests might be overly stressful and threatening. Future research should test the impact of mastery testing policies on students' motivation-related beliefs and their future academic choices, such as whether they remain enrolled in STEM majors after taking introductory chemistry courses with mastery-based testing systems. To support students' motivation, these studies could also vary the type of feedback that students receive after mastery-based tests (e.g., "wise" feedback that communicates an instructor's high expectations and the belief that a student can meet them, see Yeager et al., 2014), and they should also track the amount of time that students spend reviewing feedback, a limitation of the current study.

It will also be important to expand beyond the correlational methods used in this project and conduct experiments. These studies could recruit many instructors and randomly assign their different sections to mastery-based testing or control, although there may be better-powered, creative methods for randomizing students to different testing policies within classes (e.g., a within-subjects, crossover design in which students are randomly assigned access to mastery-based testing for different units of a course).

In conclusion, our findings highlight the potential benefits of mastery-based testing systems for promoting learning, they clarify the longitudinal process by which the course's incentive structure motivates engagement over time, and they suggest that mastery-based testing policies alone (without high-quality resources that students can use to study between attempts) are unlikely to be successful. In the future, educators and researchers should work to more fully understand the impact of mastery-based testing systems on student motivation and long-term academic outcomes, using experimental methods that allow for inference of cause and effect. By incentivizing and supporting helpful study behaviors, educators can not only enhance learning but also support students' self-regulation and well-being.

## CRediT authorship contribution statement

**Michael W. Asher:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Data curation, Conceptualization. **Joshua D. Hartman:** Writing – review & editing, Methodology, Investigation, Data curation, Conceptualization. **Mark Blaser:** Writing – review & editing. **Jack F. Eichler:** Writing – review & editing, Methodology, Conceptualization. **Paulo F. Carvalho:** Writing – review & editing, Methodology, Conceptualization.

## Author note

This research was supported in part by National Science Foundation grants DRL-2301130 to PFC and DUE-2421279 to JDH and JFE. The authors have no other competing interests to declare. https://osf.io/6pw9k.

## Data availability

All data and code for this study are openly available at https://osf.io/6pw9k.

## Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to copyedit the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compedu.2025.105387.

## Data availability

We have shard a link to an OSF repository with our data and code.

## References

Bandura, A. (1997). *Self-efficacy: The exercise of control*. Freeman.

Beilock, S. L., Kulp, C. A., Holt, L. E., & Carr, T. H. (2004). More on the fragility of performance: Choking under pressure in mathematical problem solving. *Journal of Experimental Psychology: General, 133*(4), 584–600. https://doi.org/10.1037/0096-3445.133.4.584

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64*(1), 417–444. https://doi.org/10.1146/annurev-psych-113011-143823

Carpenter, S. K., Pan, S. C., & Butler, A. C. (2022). The science of effective learning with spacing and retrieval practice. *Nature Reviews Psychology, 1*(9), 496–511. https://doi.org/10.1038/s44159-022-00089-1

Carpenter, S. K., & Toftness, A. R. (2017). The effect of prequestions on learning from video presentations. *Journal of Applied Research in Memory and Cognition, 6*(1), 104–109. https://doi.org/10.1016/j.jarmac.2016.07.014

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20*(6), 633–642. https://doi.org/10.3758/BF03202713

Carvalho, P. F., McLaughlin, E. A., & Koedinger, K. R. (2022). Varied practice testing is associated with better learning outcomes in self-regulated online learning. *Journal of Educational Psychology, 114*(8), 1723–1742. https://doi.org/10.1037/edu0000754

Cerasoli, C. P., Nicklin, J. M., & Ford, M. T. (2014). Intrinsic motivation and extrinsic incentives jointly predict performance: A 40-year meta-analysis. *Psychological Bulletin, 140*(4), 980–1008. https://doi.org/10.1037/a0035661

Chamberlin, K., Yasué, M., & Chiang, I.-C. A. (2023). The impact of grades on student motivation. *Active Learning in Higher Education, 24*(2), 109–124. https://doi.org/10.1177/14697874188l9728

Dabbagh, N., Bass, R., Bishop, M., Picciano, A. G., Sparrow, J., Costelloe, S., Cummings, K., Freeman, B., Frye, M., & Porowski, A. (2019). Using technology to support postsecondary student learning: A practice guide for college and university administrators, advisors, and faculty. *Institute of Education Sciences, what works clearinghouse, National Center for Education evaluation and regional assistance*.

Elliot, A. J., & Harackiewicz, J. M. (1996). Approach and avoidance achievement goals and intrinsic motivation: A mediational analysis. *Journal of Personality and Social Psychology, 70*(3), 461–475. https://doi.org/10.1037/0022-3514.70.3.461

Fernandez, O. E. (2021). Second chance grading: An equitable, meaningful, and easy-to-implement grading system that synergizes the research on testing for learning, mastery grading, and growth mindsets. *Primus, 31*(8), 855–868. https://doi.org/10.1080/10511970.2020.1772915

Firth, J., Rivers, I., & Boyle, J. (2021). A systematic review of interleaving as a concept learning strategy. *The Review of Education, 9*(2), 642–684. https://doi.org/10.1002/rev3.3266

Fyfe, E. R., De Leeuw, J. R., Carvalho, P. F., Goldstone, R. L., Sherman, J., Admiraal, D., Alford, L. K., Bonner, A., Brassil, C. E., Brooks, C. A., Carbonetto, T., Chang, S. H., Cruz, L., Czymoniewicz-Klippel, M., Daniel, F., Driessen, M., Habashy, N., Hanson-Bradley, C. L., Hirt, E. R., … Motz, B. A. (2021). ManyClasses 1: Assessing the generalizable effect of immediate feedback versus delayed feedback across many college classes. *Advances in Methods and Practices in Psychological Science, 4*(3), Article 25152459211027575. https://doi.org/10.1177/25152459211027575

Hartman, J. D., & Eichler, J. F. (2024). Implementing a mastery grading in large enrollment general chemistry: Improving outcomes and reducing equity gaps. *Education Sciences, 14*(11), Article 1224. https://doi.org/10.3390/educsci14111224

Hidi, S., & Harackiewicz, J. M. (2000). Motivating the academically unmotivated: A critical issue for the 21st century. *Review of Educational Research, 70*(2), 151–179. https://doi.org/10.3102/00346543070002151

Howitz, W. J., McKnelly, K. J., & Link, R. D. (2021). Developing and implementing a specifications grading system in an organic chemistry laboratory course. *Journal of Chemical Education, 98*(2), 385–394. https://doi.org/10.1021/acs.jchemed.0c00450

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*(2), 75–86. https://doi.org/10.1207/s15326985ep4102_1

Koedinger, K. R., Carvalho, P. F., Liu, R., & McLaughlin, E. A. (2023). An astonishing regularity in student learning rate. *Proceedings of the National Academy of Sciences, 120*(13), Article e2221311120. https://doi.org/10.1073/pnas.2221311120

Koedinger, K. R., Kim, J., Jia, J. Z., McLaughlin, E. A., & Bier, N. L. (2015). Learning is not a spectator sport: Doing is better than watching for learning from a MOOC. *Proceedings of the Second (2015) ACM Conference on Learning @ Scale*. https://doi.org/10.1145/2724660.2724681

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(4), 989–998. https://doi.org/10.1037/a0015729

Kroeper, K. M., Muenks, K., Canning, E. A., & Murphy, M. C. (2022). An exploratory study of the behaviors that communicate perceived instructor mindset beliefs in college STEM classrooms. *Teaching and Teacher Education, 114*, Article 103717. https://doi.org/10.1016/j.tate.2022.103717

Kulik, J. A., & Kulik, C.-L. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research, 58*(1), 79. https://doi.org/10.2307/1170349

Lovett, M., Meyer, O., & Thille, C. (2008). Measuring the effectiveness of the OLI statistics course in accelerating student learning. *Journal of Interactive Media in Education, 2008*(1), 1–16. https://doi.org/10.5334/2008–14

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology, 103*(2), 399–414. https://doi.org/10.1037/a0021782

Morphew, J. W., Silva, M., Herman, G., & West, M. (2020). Frequent mastery testing with second-chance exams leads to enhanced student learning in undergraduate engineering. *Applied Cognitive Psychology, 34*(1), 168–181. https://doi.org/10.1002/acp.3605

Nichols, L., & Islas, Á. (2016). Pushing and pulling emerging adults through college: College generational status and the influence of parents and others in the first year. *Journal of Adolescent Research, 31*(1), 59–95. https://doi.org/10.1177/0743558415586255

Pan, S. C., & Carpenter, S. K. (2023). Prequestioning and pretesting effects: A review of empirical research, theoretical perspectives, and implications for educational practice. *Educational Psychology Review, 35*(4), 97. https://doi.org/10.1007/s10648-023-09814-5

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin, 144*(7), 710–756. https://doi.org/10.1037/bul0000151

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology, 8*, 422. https://doi.org/10.3389/fpsyg.2017.00422

Penk, C., & Richter, D. (2017). Change in test-taking motivation and its relationship to test performance in low-stakes assessments. *Educational Assessment, Evaluation and Accountability, 29*(1), 55–79. https://doi.org/10.1007/s11092-016-9248-7

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers, 36*(4), 717–731. https://doi.org/10.3758/BF03206553

Pulfrey, C., Buchs, C., & Butera, F. (2011). Why grades engender performance-avoidance goals: The mediating role of autonomous motivation. *Journal of Educational Psychology, 103*(3), 683–700. https://doi.org/10.1037/a0023911

Rawson, K. A., Vaughn, K. E., Walsh, M., & Dunlosky, J. (2018). Investigating and explaining the effects of successive relearning on long-term retention. *Journal of Experimental Psychology: Applied, 24*(1), 57–71. https://doi.org/10.1037/xap0000146

Ring, J. (2017). ConfChem conference on select 2016 BCCE presentations: Specifications grading in the flipped organic classroom. *Journal of Chemical Education, 94*(12), 2005–2006. https://doi.org/10.1021/acs.jchemed.6b01000

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

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. https://doi.org/10.1037/a0037559

Ryan, R. M., & Weinstein, N. (2009). Undermining quality teaching and learning: A self-determination theory perspective on high-stakes testing. *Theory and Research in Education, 7*(2), 224–233. https://doi.org/10.1177/1477878509104327

Schooler, L. J., & Anderson, J. R. (1990). The disruptive potential of immediate feedback. *Proceedings of the twelfth annual conference of the cognitive science Society*.

Senko, C., Hulleman, C. S., & Harackiewicz, J. M. (2011). Achievement goal theory at the crossroads: Old controversies, current challenges, and new directions. *Educational Psychologist, 46*(1), 26–47. https://doi.org/10.1080/00461520.2011.538646

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52*(6), 613–629. https://doi.org/10.1037/0003-066X.52.6.613

Von Der Embse, N., Jester, D., Roy, D., & Post, J. (2018). Test anxiety effects, predictors, and correlates: A 30-year meta-analytic review. *Journal of Affective Disorders, 227*, 483–493. https://doi.org/10.1016/j.jad.2017.11.048

Warburton, E. C., Burgarin, R., & Nuñez, A.-M. (2001). *Bridging the gap: Academic preparation and postsecondary success of first-generation students* (No. NCES 2001–153). *National Center for Education Statistics*. http://nces.ed.gov/pubs2001/2001153.pdf.

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin, 147*(4), 399–435. https://doi.org/10.1037/bul0000309

Yeager, D. S., Purdie-Vaughns, V., Garcia, J., Apfel, N., Brzustoski, P., Master, A., Hessert, W. T., Williams, M. E., & Cohen, G. L. (2014). Breaking the cycle of mistrust: Wise interventions to provide critical feedback across the racial divide. *Journal of Experimental Psychology: General, 143*(2), 804–824. https://doi.org/10.1037/a0033906