7-30-2021

# Persistent Miscalibration for Low and High Achievers Despite Practice Test Feedback in an Introductory Biology Course

Jennifer L. Osterhage
*University of Kentucky*, jennifer.osterhage@uky.edu

# Persistent Miscalibration for Low and High Achievers Despite Practice Test Feedback in an Introductory Biology Course

Digital Object Identifier (DOI)
https://doi.org/10.1128/jmbe.00139-21

Notes/Citation Information
Published in *Journal of Microbiology & Biology Education*, v. 22, no. 2.

Research Article

JMBE Journal of Microbiology & Biology Education

# Persistent Miscalibration for Low and High Achievers despite Practice Test Feedback in an Introductory Biology Course

Jennifer L. Osterhage[a]

[a]Department of Biology, University of Kentucky, Lexington, Kentucky, USA

**Students' ability to accurately judge their knowledge is crucial for effective learning. However, students' perception of their current knowledge is often misaligned with their actual performance. The relationship between learners' perception of their performance and their actual performance on a task is defined as calibration. Previous studies have shown significant student miscalibration in an introductory biology course: students' predicted exam scores were, on average, significantly higher than their actual scores. The goal of this study was to determine whether completion of a practice test before exams would result in better performance and calibration. The hypothesis was that students who completed a practice test would perform better and be better predictors of their performance on exams than students who did not engage in practice testing. As predicted, students who voluntarily completed a practice test, on average, performed better and were more calibrated than students who did not. Importantly, however, many of the lowest-performing students continued to significantly overestimate their knowledge, predicting higher scores on the exam than they actually earned, despite feedback from practice tests. In contrast, practice testing was associated with underconfidence in high-performing students. These findings indicate that practice tests may enhance calibration for many students. However, additional interventions may be required for the lowest-performing students to become better predictors of their performance.**

**KEYWORDS** calibration, metacognition, Dunning-Kruger effect, testing effect, undergraduate biology

## INTRODUCTION

Almost all undergraduate students enter introductory courses expecting to earn an "A" or "B" grade (1). However, gateway undergraduate science courses tend to have high failure (D/F) and withdrawal (W) rates (2), indicating that many students earn substantially lower grades than they initially anticipated. This incongruence has important consequences: science students may change their majors to a nonscience field because their grades in science courses did not match their expectations (3).

Metacognition, simply defined as the ability to think about one's own thinking (4), consists of two key elements: metacognitive knowledge and metacognitive regulation (5). Metacognitive knowledge includes learning processes, awareness of effective learning strategies, and the ability to distinguish between knowing and not knowing. Metacognitive regulation refers to learners' ability to accurately evaluate strengths and weaknesses, reflect on the success of their strategies, and adjust accordingly. Metacognitively aware students accurately assess a task, make plans, and effectively self-monitor during learning (6). For metacognitively unaware students, the perception of their current knowledge is often misaligned with their performance. When students do not grasp the limits of their understanding, they are at risk of underperformance and academic failure (7). Instructor strategies to promote student metacognition have shown promise, but intervention studies have lagged behind foundational research in this area (8).

### Calibration and the Dunning-Kruger effect

Calibration occurs when learners' judgments are closely related to their actual performance on a task (9). Calibration measures have been used as indicators of metacognitive monitoring ability. There are multiple methods used to assess calibration, including calculation of the difference between predictions of performance and actual performance (10). Both over- and underestimates lead to miscalibration, the inaccuracy in judgment between perception of performance and actual performance (11). Judgment

**Citation** Osterhage JL. 2021. Persistent miscalibration for low and high achievers despite practice test feedback in an introductory biology course. J Microbiol Biol Educ 22:e00139-21. https://doi.org/10.1128/jmbe.00139-21.

Address correspondence to Department of Biology, University of Kentucky, Lexington, Kentucky, USA. E-mail: Jennifer.osterhage@uky.edu.

Received: 8 March 2021, Accepted: 10 June 2021,
Published: 30 July 2021

errors may influence study efforts, resulting in lower academic success (12). For example, overconfident students may prematurely cease study efforts when they believe that they have mastered concepts (13). Importantly, students who are poorly calibrated are more likely to earn lower course grades than calibrated students (14).

The least competent individuals are the most likely to be overconfident in their performance judgments. This cognitive bias, in which unskilled individuals are the most likely to overestimate their ability, is named the Dunning-Kruger effect (15). The Dunning-Kruger effect has been observed in multiple studies across various contexts (16–20) and has been shown to persist even when learners are presented with accurate information about their skill level (21–23). The Dunning-Kruger effect has been attributed to metacognitive differences between groups of learners. While skilled individuals are metacognitively aware, unskilled individuals have gaps or distortions in their knowledge that do not allow them to realize how unskilled they are (15).

Inaccurate performance judgments may persist as a self-protective mechanism. A qualitative analysis (24) observed distinct patterns of thinking for both high- and low-achieving students. High-achieving students reported underestimating themselves as to not appear immodest and to avoid disappointment. These students also used underconfidence as a motivational strategy to stimulate study efforts. Overconfident students, in contrast, were motivated by optimistic predictions of their performance. In agreement with this analysis, Helzer and Dunning (25) found that overconfident individuals gave more weight to their aspirations than evidence of past achievement when making performance judgments. Taken together, these studies suggest that both high- and low-achievement individuals use subjective measures rather than objective information to inform their performance judgments.

Overconfidence leads to the premature termination of studying and lower levels of retention (13). Given the high-stakes nature of exams in many undergraduate science courses, accurate judgments of preparedness for summative assessments are particularly important in these contexts. Therefore, inaccurate self-evaluation poses a significant risk for low grades in college science courses. In support of this assertion, in a first-semester chemistry course, the extent of overconfidence on a pretest predicted the likelihood of a failing final grade (26). In an introductory biology course, lower-achieving students had the most inaccurate estimates of their performance (27). In other studies of introductory biology and chemistry students, the lowest-performing students overestimated their exam performance, overall grades, and class rank (19, 28–30). These studies demonstrate that overconfidence is associated with lower course grades in introductory science courses, underscoring the important role of metacognitive awareness in these contexts.

## The testing effect (retrieval practice)

Taking tests during the learning process has been shown to lead to better long-term retention than other study methods such as rereading. The finding that testing is superior to restudying, called the testing effect or retrieval practice effect, is supported by robust empirical evidence (31–35). The contribution to long-term memory is not the only benefit of testing during learning: tests can also serve as a monitoring tool by giving students feedback about their level of understanding (36). Based on these findings, the use of testing as a learning strategy has been encouraged (31–37).

The use of testing for learning by students in undergraduate science courses has been described. In an undergraduate chemistry course, retrieval practice strategies were not as widely used as review-type strategies (38). In contrast, answering questions from old exams was the most popular study strategy for students in an introductory biology course (39). Students earning a "D" or "F" on the first exam in an introductory biology course reported lower usage of self-testing than higher-achieving students (40). Another study showed that self-testing increased over time among students in an introductory chemistry course (41). These studies highlight the importance of retrieval practice in college science courses.

## Retrieval practice and calibration

Because more information is available to inform performance judgments, it has been asserted that retrieval practice activities, followed by feedback about one's performance, should enhance calibration (42). However, previous studies investigating the relationship between retrieval practice activities and calibration have yielded mixed results. Some studies support the assertion that learners have more accurate judgments after retrieval practice (43–45). However, practice testing was associated with greater miscalibration in other studies. For example, graduate students in a research methods course who completed practice tests were worse predictors of exam performance than students who did not (46). In another study, learners became increasingly underconfident with more practice across a variety of contexts (47). This effect was termed the underconfidence-with-practice (UWP) effect and has since been observed in multiple contexts (48–50). The conflicting reports about the nature of the relationship between practice testing and calibration suggest that feedback from practice has inconsistent effects on learners with different characteristics. Relatively little is known about how feedback from testing may differentially impact calibration for students with different levels of achievement.

Given the importance of accurate judgments of learning for academic success and the known benefits of testing, feedback from practice tests could be a particularly powerful tool to enhance success in science courses. The purpose of this study was to compare the predicted and actual exam performances of students who received feedback from a practice test with those of students who did not. The specific research questions were as follows:

1. Do practice tests enhance performance and calibration in an introductory biology course?
2. Do practice tests differentially affect calibration based on achievement level?

The hypothesis was that, on average, students who complete a practice test would perform better and be better predictors of their performance on exams than students who did not complete a practice test. Given previous research on the Dunning-Kruger and UWP effects, it was hypothesized that practice tests may have distinct effects based on achievement level. Specifically, the prediction was that low-achieving students would continue to be overconfident and that high-achieving students would become more underconfident after feedback from practice testing.

## METHODS

### Participants

Study participants were enrolled in Introductory Biology I at a 4-year public institution in the southeastern United States. Consenting participants ($n = 341$) were enrolled in one of four course sections during the Fall 2019 semester. All participants completed each exam in the course.

### Course description and setting

Introductory Biology I is a required course for the undergraduate biology major and other science and prehealth majors across the university. Contact hours consisted of 150 minutes per week throughout a 16-week semester. Topics covered included the nature of science, evolution, gene expression, cell division, inheritance, ecology, and biodiversity.

In the study semester, each of the four sections was taught by a different instructor. J. L. Osterhage was the instructor for one section. Learning outcomes, grading schemes, homework questions, practice tests, and actual exams were uniform across sections. The course schedule was the same across sections except for adjustments for class meeting patterns. All students were provided with the same course packet, which included section summaries, learning outcomes, practice questions, and study checklists. Instructors used different notes and practice questions during class time. Activities to promote self-evaluation were embedded throughout all sections: (i) clicker questions, which allowed students to see the percentage of classmates who chose each answer; (ii) group quizzing, in which students were encouraged to consider how many of their answers they changed after group discussion; and (iii) the availability of additional practice questions in the learning management system (LMS) and the course packet. Effective study strategies were discussed in class and included in the course packet. In all sections, students were encouraged, but not required, to complete the online practice tests described below. All instructors discussed how feedback from the practice tests should be used to identify areas of strength and weakness before the exam.

### Design and procedures. (i) Online practice tests

All students were provided with the same online practice test for each exam in the course LMS. Practice tests consisted of 45 to 60 multiple-choice questions used on exams from previous semesters, which were chosen to accurately represent the content of exams in the study semester. Except for the cumulative practice test, which contained 60 questions (compared to 86 questions on the final exam), the number of practice test questions matched the number of questions on the exam. The time limit of practice tests 1 to 3 was 75 min. The time limit for the final practice test was 120 min. In each case, the practice test time limit matched the time frame given for the exam. Practice tests were not proctored. Completion of the practice test was voluntary and did not affect the course grade. Correct answers were not visible until students submitted their answers. After submitting the practice test, students immediately received their score and could view correct and incorrect answers. No other feedback (e.g., explanations of answer choices) was provided.

### (ii) Exams

Exams were the same across sections and were completed during a common hour exam period. Students received a paper copy of the exam and filled in their answers on a Scantron sheet. The first and third exams consisted of 50 multiple-choice questions. The second exam consisted of 45 multiple-choice questions. The final exam was partially cumulative and consisted of 86 multiple-choice questions. No questions were duplicated exactly between the practice test and the actual exam, but there was considerable overlap between question styles and concepts tested.

The cover sheet of the exam included the statement, "Enter the numerical percentage (0–100) that you expect to earn on this exam _____." Directly before beginning the exam, students were prompted to fill in the blank. Predictions of exam scores were gathered to capture students' perceived levels of preparedness for exams before they occurred. After each exam, exam questions and answer keys were posted on the LMS.

### (iii) Data sources and analysis

Data collection and analysis were approved by the university's institutional review board (IRB) (approval number 53301). Data were analyzed after final course grades were submitted for all students. A multiple-linear-regression model that included course instructor as a random effect was not significant ($P = 0.9465$) (data not shown). Therefore, data from all sections were combined for analysis.

All consenting students were included in the calculation of the performance quartile for each exam. Students who did not enter a predicted exam score were excluded from the calculation of discrepancy scores.

TABLE 1
Percentages of students completing practice tests

| Group (practice test score [%]) | % of students who completed the practice test | | | |
|---|---|---|---|---|
| | Exam 1 | Exam 2 | Exam 3 | Final exam |
| A (≤20 or no practice test) | 18.5 | 12.8 | 16.8 | 45.4 |
| B (>20) | 81.5 | 87.2 | 83.2 | 54.6 |

Discrepancy scores were calculated as the difference between students' predicted and actual percentage scores on each exam (predicted score minus actual score). Positive raw discrepancy scores indicated that students overestimated their performance. Negative raw scores indicated that students underestimated their performance.

Student scores on the practice test were obtained from the LMS. Students were grouped into two categories: those who did not complete the practice test or scored ≤20% (the score expected for random guessing) (group A) and those who scored >20% (group B) on the practice test. The 20% cutoff was chosen because random guessing was not expected to have the same benefits for learning as an earnest effort. This approach was supported by one-way analysis of variance (ANOVA), which indicated no significant difference in exam performance and calibration between students who did not complete the practice test and those who scored ≤20%.

Mean discrepancy scores and mean exam scores were calculated for group A and group B and analyzed for each exam using Student's *t* test. For group B, practice test scores were plotted against actual and expected exam scores. $R^2$ was calculated for each scatterplot, and the relative strengths of correlations were compared using Fisher *r*-to-*z* transformation. Expected and actual scores were plotted against the actual percentile rank for each group. The area between the best-fit curves was highlighted to identify the relative contributions of each quartile to the overall discrepancy scores. Mean discrepancy scores were calculated separately for the lowest quartile and the highest quartile. To investigate the differences among means, data were analyzed using two-way ANOVA with the independent variables exam number and group and by three-way ANOVA including quartile as an additional variable. Pairwise comparisons of the means were performed *post hoc* using Tukey's multiple-comparison tests. For each exam, Cohen's *d* was calculated to measure the effect size. Effect sizes were averaged across exams to determine the mean effect size.

## RESULTS

### Do practice tests enhance performance and calibration in an introductory biology course?

The majority of students completed the practice test for exams 1 to 3, while fewer students completed the final practice test (Table 1). Consistent with the testing effect, students who completed the practice test (group B) earned significantly higher exam scores for exams 1, 3, and 4 than students who did not (group A) (Fig. 1, top, and Table 2). For exam 2, the difference in performances between groups A and B was not statistically significant. The effect size of practice testing on performance was greatest for the first exam (0.48), with a mean effect size across exams of 0.37.

The degree to which students' predicted scores were calibrated with their actual scores on exams was measured by calculating a discrepancy score, defined as the difference between students' earned exam percentage and the percentage predicted before taking the exam. For exam 1, the mean discrepancy score for students who either did not complete the practice test or scored lower than 20% (group A) (mean = 10.0) was significantly higher than that for students who completed the practice test (group B) (mean = 6.6; effect size = 0.31) (Fig. 1, bottom left). The differences in discrepancy scores between student groups were not statistically significant for exams 2 to 4 (Fig. 1, bottom). Across exams, the mean effect size of practice tests on calibration was 0.21. For both groups A and B, discrepancy scores decreased as the semester progressed (compare the scales in Fig. 1, bottom).

Practice test scores were correlated with earned exam scores (Fig. 2, right), indicating that practice tests provided accurate feedback about the level of preparedness for exams. The correlation between student predictions of exam performance and actual performance, however, was less robust (Fig. 2, compare left and right panels). Even as average miscalibration decreased as the semester progressed, the correlation between practice test scores and expected scores did not change, indicating that feedback from the practice test was not the only factor that students used when making performance predictions.

### Do practice tests differentially affect calibration based on achievement level?

Best-fit lines of actual and predicted exam scores were graphed against the percentile rank of performance. The Dunning-Kruger effect was observed whether or not students completed the practice test: the lowest-performing students were least calibrated when predicting their scores (Fig. 3, dark red). The correlation between actual and predicted exam scores increased as the semester progressed (Fig. 3, compare dark red areas among exams 1 to 4).

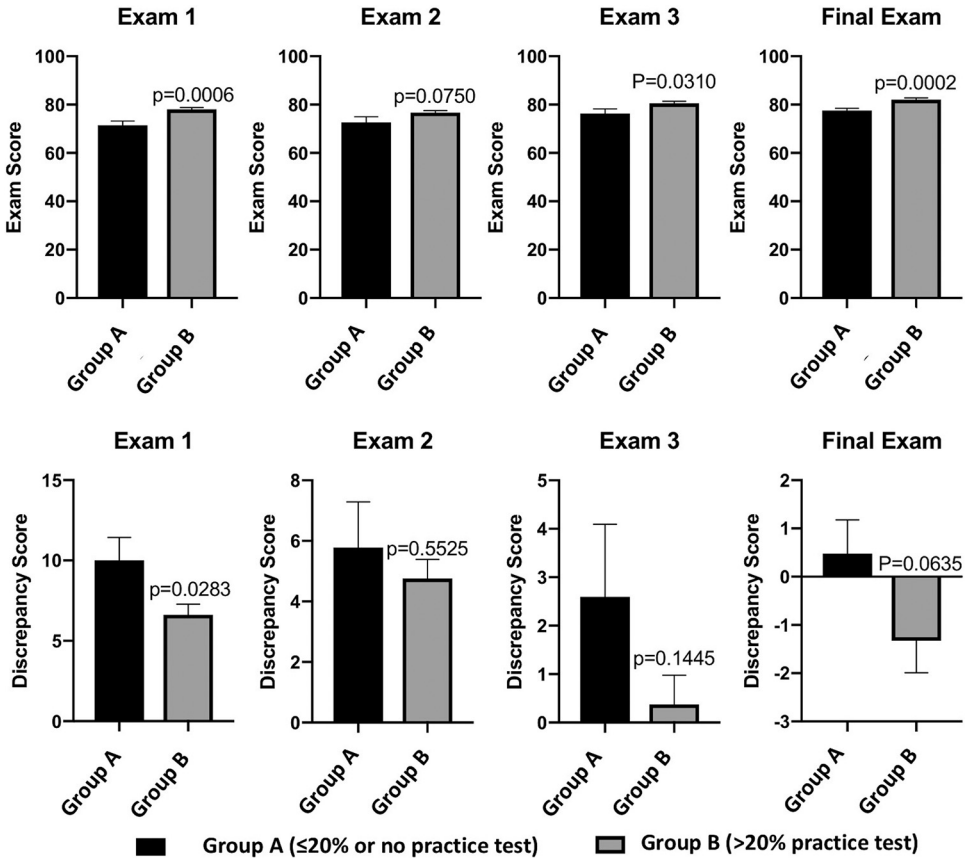


FIG 1. Completion of practice tests associated with improved exam performance and lower discrepancy scores. (Top) Students who completed the online practice test had significantly higher exam scores for exam 1, exam 3, and the final exam, with a trend toward improved performance on exam 2 (by a *t* test) ($n = 341$). (Bottom) Completion of the online practice exam with a score of over 20% (group B) was associated with significantly reduced discrepancies between the expected and actual earned scores of exam 1 compared with group A (by a *t* test) ($n = 341$). The trend was similar, although not statistically significant, for exam 2, exam 3, and the final exam.

As a group, the lowest-performing students became less overconfident as the semester progressed (Fig. 4a, compare bar heights for exams 1 to 4). However, practice testing did not lower miscalibration for low-performing students (Fig. 4a, compare black and white bars). For the highest quartile of students, completion of the practice test was associated with greater underestimation of actual exam performance when all exams were analyzed together ($P = 0.05$; effect size = $-0.34$).

When grouping the lowest and highest quartiles of students across all exams, those who completed practice tests were significantly more miscalibrated than those who did not complete practice tests ($P = 0.02$). Taking the practice test exaggerated the tendency of both low- and high-achieving students to be miscalibrated albeit in different directions. Practice test completion did not lower overconfidence in low-achieving students and increased underconfidence in high-achieving students.

## DISCUSSION

Achievement in introductory science courses is negatively affected by miscalibration, a mismatch between performance judgment and actual performance. The central goal of this study was to determine whether feedback from practice

TABLE 2
Mean exam percentages

| Group (practice test score [%]) | Mean exam score (%) | | | |
|---|---|---|---|---|
| | Exam 1 | Exam 2 | Exam 3 | Final exam |
| A (≤20 or no practice test) | 71.4 | 72.7 | 76.3 | 77.6 |
| B (>20) | 78.0 | 76.7 | 80.6 | 82.0 |

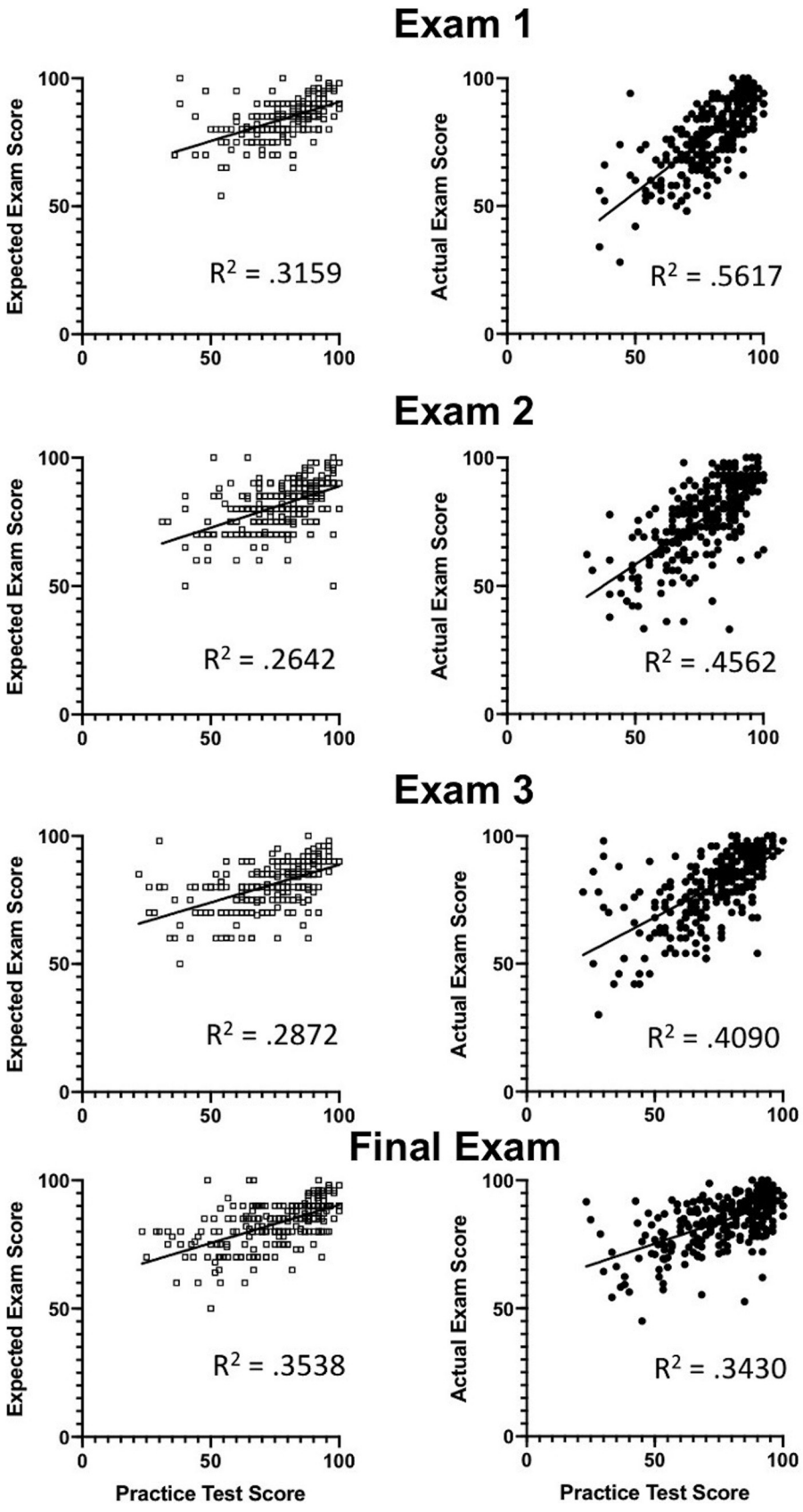


FIG 2. The practice test score is more highly correlated with the actual exam score than the expected score. For group B, practice test scores were plotted against expected exam scores (left) and actual exam scores (right). $R^2$ was calculated for each scatterplot, and relative strengths of correlations were compared using Fisher *r*-to-*z* transformation.

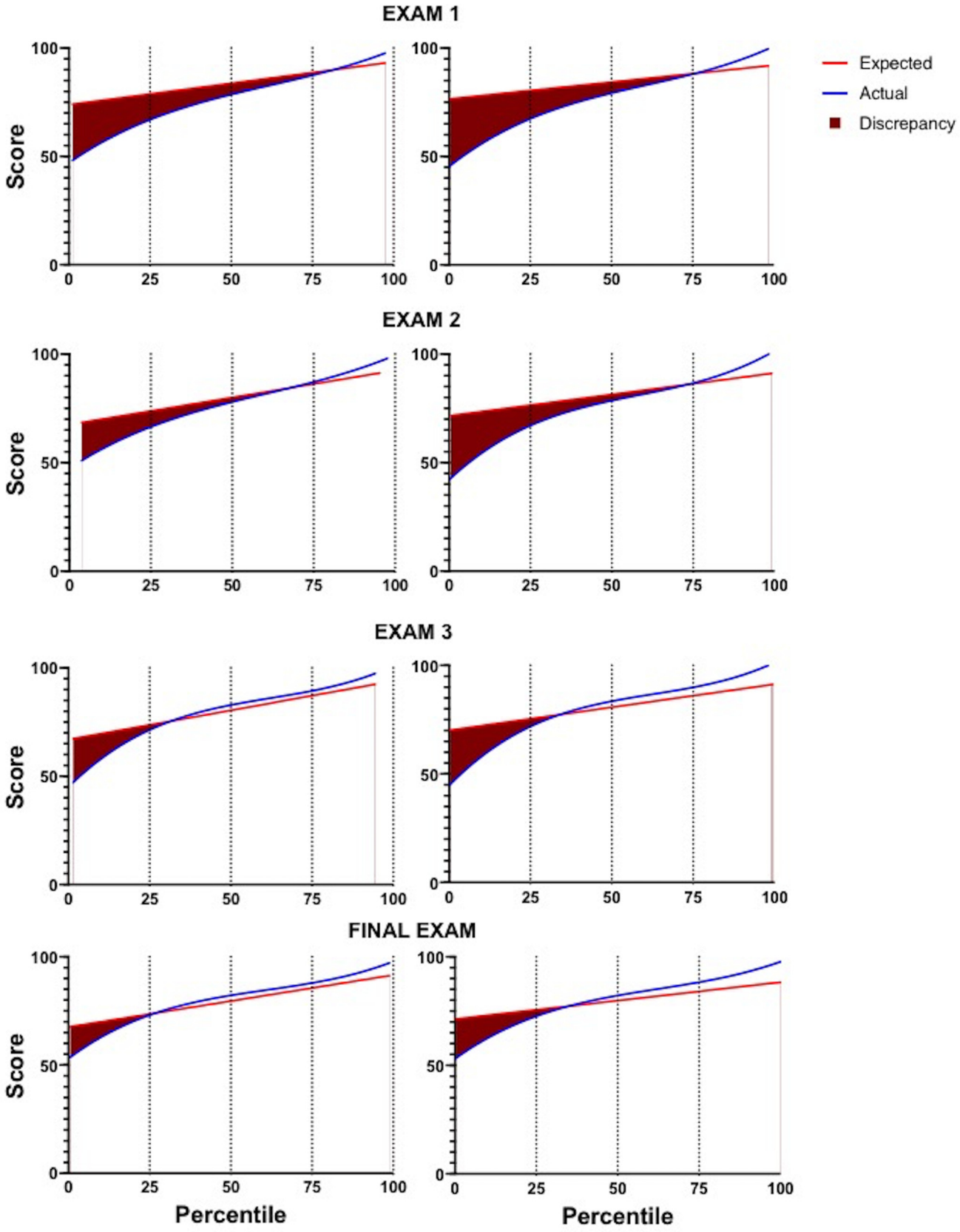


FIG 3. Distribution of expected scores and actual scores by percentile rank. Best-fit lines of predicted and actual scores graphed by percentile rank of actual scores were plotted for each exam. Expected scores and actual scores are depicted. The area between curves (dark red) represents overestimation. The majority of the discrepancy between actual and expected scores can be attributed to the lowest quartile.

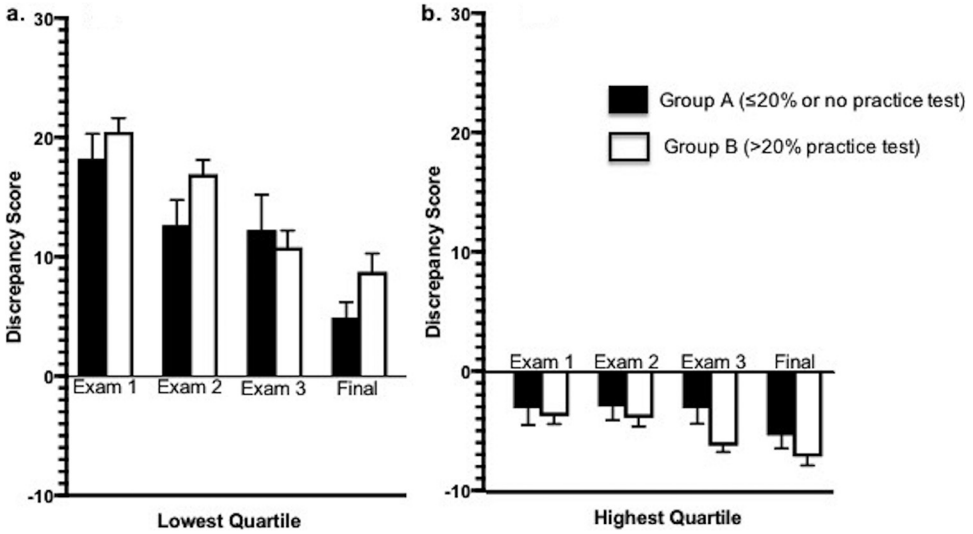


FIG 4. Practice exams do not lower miscalibration for the highest- and lowest-performing students. Discrepancy scores were plotted for the lowest quartile (a) and the highest quartile (b) based on exam score. (a) The lowest-performing students who completed the practice test trended toward more miscalibration than students who did not complete the practice test. (b) For the highest quartile, completion of the practice test was associated with underestimation of performance.

testing improved performance and calibration of students in an introductory biology course. The data presented here are consistent with the testing effect, in which students who test themselves as a study strategy perform better on summative assessments. This finding was in agreement with a vast body of research on the benefits of testing for learning (for a review, see reference 37).

The relationship between practice testing and calibration was more complex. Early in the semester, students who completed the practice test were, on average, less miscalibrated than students who did not complete it. These results indicate that feedback, both from practice tests and from prior performance, contributed to enhanced calibration. While the overall impact of practice testing was better calibration coinciding with higher exam scores, a closer look at the distribution of expectations and performances revealed diverging effects on the highest and lowest quartiles. Practice testing did not mitigate the Dunning-Kruger effect, with many of the lowest-performing students continuing to predict much higher exam scores relative to their performance on the practice test (Fig. 2, left). The miscalibration of low-achieving students decreased throughout the semester, mainly due to better performance on exams, suggesting that practice testing did not significantly influence performance predictions. In contrast, high-achieving students underestimated their knowledge after practice testing and increasingly underpredicted their performance as the semester progressed. This may be the first study that has revealed a trend toward greater miscalibration for low- and high-performing students who complete practice tests than for those who do not.

Low-achieving students may use global self-concepts of academic ability rather than objective feedback to inform performance estimates (17). If that is the case, these students may require more feedback over a longer period to adjust their self-concepts. Reasoning ability is correlated with achievement in introductory biology (51). Low reasoning ability may affect both biology achievement and calibration because the skills and knowledge required to be successful in a discipline are the same ones required to assess one's level of understanding (15).

Previous studies have demonstrated that the more practice an individual engages in, the more underconfident they become (the UWP effect). This study provides support for this effect, especially among high-performing students. The highest-performing students became increasingly underconfident relative to their abilities as the semester progressed, which was exacerbated by the completion of practice tests.

Many factors could contribute to the distinct effects of practice testing on calibration based on achievement level. The level of similarity between the practice test and the exam may differentially affect students based on their skill level. If the exam included many items that were not represented on the practice test, calibration could be negatively affected, especially for students with low understanding. In support of this hypothesis, students in an introductory biology course performed worse on new test items than on items that they were familiar with from old exams (39). In addition, a recent study demonstrated that performance on practice-tested items was significantly greater than that on nontested items (45). Even though students in this study were informed that the questions on practice tests would not be duplicated on exams, lower-performing students may have tried to memorize the answers to practice test questions and used their success at memorization as a basis for their predictions.

While not measured in this study, the amount and timing of retrieval practice activities may vary significantly

between students. For example, a previous study indicated that the majority of students mass their studying the evening before an exam, which limits the use of effective study strategies (52). It is possible that students completed the practice test too close to the exam to affect their study strategies and calibration. Learners also tend to test themselves only under conditions that encourage retrieval success (36). It is possible that the lowest-performing students in the class engaged in fewer of the available alternative practice activities because they were not confident that they could be successful.

In agreement with a previous publication (27), in this study, average miscalibration decreased as the semester progressed. This finding suggests that feedback from prior exam performance informed predicted performance for later exams. The precise contributions of feedback from prior exam performance and that from practice testing were not measured. However, the significant difference in miscalibration between students who did and those who did not complete the practice test before exam 1 (Fig. 1) suggests that practice testing may be particularly important to mitigate miscalibration early in the semester, before other forms of feedback are available.

### Limitations and future directions

This study was limited in a few ways. First, it is possible that the learner characteristics of those who completed practice tests differed from those of students who did not. For example, students who completed the practice tests may have been more likely to believe that they could be successful (36). Students who completed practice tests could have had higher levels of metacognitive knowledge and regulation than others. Previous studies have shown that prior knowledge affects calibration when exam items were not previously tested by retrieval practice (45). In this study, prior knowledge was not measured, so it is not clear whether this may have affected calibration and exam performance. In addition, study strategy usage other than the practice test was not monitored in this study. It is possible that study strategies varied across the semester. For example, all previous exams were available for students to prepare for the final exam. The availability of these exams may explain why fewer students completed the practice test to prepare for the final exam. While no questions were exactly duplicated between the practice test and the actual exam, there were many questions on the practice test that required types of application and analysis similar to those required for the exam. This study did not track how many of the items on the practice test were directly comparable to exam questions. In addition, the timing of practice test completion and the amount of time spent engaging with the practice test were not tracked in this study. It is possible that some students took the practice test too close to the exam or did not engage with the practice test fully enough to reap the metacognitive benefits. In this study, group A was a heterogeneous population, consisting of students who did not open the practice test and those who earned a score of <20%. While it is not expected that random guessing or leaving most questions unanswered would confer cognitive or metacognitive benefits, these students did have access to the correct answers on the practice test. Therefore, it is possible that group A consisted of two fundamentally different subgroups. However, the level of engagement with the answer key could not be measured to identify any potential subpopulations.

In the present study, students were asked to predict their scores on exams directly before taking them as an indicator of their perceived level of preparedness. However, we cannot rule out that some students changed their predictions after taking the exam (postdiction). Future studies could compare prediction and postdiction performance estimates to determine if exam completion affects performance judgments.

In this study, feedback from the practice test consisted of the score earned and the ability to view correct and incorrect answers. It would be interesting to determine whether additional feedback, such as narrative descriptions of answer choices, would enhance the benefits of testing as a learning strategy. It would also be interesting to determine if these findings would apply to assessments other than multiple-choice-based exams. Future studies could also explore whether training about how to use practice testing as a study tool would enhance calibration and exam performance.

### Implications for instructors

The findings presented here suggest several instructor practices. First, providing full-length practice tests as a form of formative assessment may enhance overall student performance and calibration. Utilization of practice testing is particularly important for calibration early in the semester. Students may also benefit from explicit descriptions of the purpose of formative assessments and how to utilize the feedback. Instructors should keep in mind, however, that the use and effectiveness of this strategy will vary among students. Low-performing students may be resistant to changing their self-views and may require additional interventions to become better calibrated. If summative assessments represent a large portion of the students' final grades, it will be especially important to reach miscalibrated students early in the semester, ideally before the first summative assessment. These strategies add to the existing toolkit (8) that instructors can utilize to foster student metacognition and enhance learning in undergraduate science courses.

## ACKNOWLEDGMENTS

I thank Katie Meikel and Nicholas Strobl for data entry. I thank Ellen Usher, Peter Mirabito, Ann Morris, and Arnold Stromberg for their productive feedback.

Research reported in this publication was supported by an institutional development award (IDeA) from the National

Institute of General Medical Sciences of the National Institutes of Health under grant number P20GM103436.

I have no conflicts of interest to declare.

## REFERENCES

1. Beattie G, Laliberté J-WP, Oreopoulos P. 2018. Thrivers and divers: using non-academic measures to predict college success and failure. Econ Educ Rev 62:170–182. https://doi.org/10.1016/j.econedurev.2017.09.008.
2. Freeman S, Haak D, Wenderoth MP. 2011. Increased course structure improves performance in introductory biology. CBE Life Sci Educ 10:175–186. https://doi.org/10.1187/cbe.10-08-0105.
3. Stinebrickner R, Stinebrickner TR. 2014. A major in science? Initial beliefs and final outcomes for college major and dropout. Rev Econ Stud 81:426–472. https://doi.org/10.1093/restud/rdt025.
4. Mahdavi M. 2014. An overview: metacognition in education. Int J Multidiscip Curr Res 2:529–535.
5. Flavell JH. 1979. Metacognition and cognitive monitoring: a new area of cognitive-developmental inquiry. Am Psychol 34:906–911. https://doi.org/10.1037/0003-066X.34.10.906.
6. Ambrose SA, Bridges MW, DiPietro M, Lovett MC, Norman MK. 2010. How learning works: seven research-based principles for smart teaching. Jossey-Bass, San Francisco, CA.
7. Serra MJ, DeMarree KG. 2016. Unskilled and unaware in the classroom: college students' desired grades predict their biased grade predictions. Mem Cognit 44:1127–1137. https://doi.org/10.3758/s13421-016-0624-9.
8. Stanton J, Sebesta A, Dunlosky J. 2021. Fostering metacognition to support student learning and performance. CBE Life Sci Educ 20:fe3. https://doi.org/10.1187/cbe.20-12-0289.
9. Hattie J. 2013. Calibration and confidence: where to next? Learn Instr 24:62–66. https://doi.org/10.1016/j.learninstruc.2012.05.009.
10. Dunlosky J, Thiede KW. 2013. Four cornerstones of calibration research: why understanding students' judgments can improve their achievement. Learn Instr 24:58–61. https://doi.org/10.1016/j.learninstruc.2012.05.002.
11. Huff JD, Nietfeld JL. 2009. Using strategy instruction and confidence judgment to improve metacognitive monitoring. Metacogn Learn 4:161–176. https://doi.org/10.1007/s11409-009-9042-8.
12. Bembenutty H. 2009. Three essential components of college teaching: achievement calibration, self-efficacy, and self-regulation. Coll Stud J 43:562–570.
13. Dunlosky J, Rawson KA. 2012. Overconfidence produces underachievement: inaccurate self evaluations undermine students' learning and retention. Learn Instr 22:271–280. https://doi.org/10.1016/j.learninstruc.2011.08.003.
14. Garavalia LS, Gredler ME. 2002. An exploratory study of academic goal setting, achievement calibration and self-regulated learning. J Educ Psychol 29:221–230.
15. Kruger J, Dunning D. 1999. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. J Pers Soc Psychol 77:1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121.
16. Caputo D, Dunning D. 2005. What you don't know: the role played by errors of omission in imperfect self-assessments. J Exp Soc Psychol 41:488–505. https://doi.org/10.1016/j.jesp.2004.09.006.
17. Ehrlinger J, Dunning D. 2003. How chronic self-views influence (and potentially mislead) estimates of performance. J Pers Soc Psychol 84:5–17. https://doi.org/10.1037/0022-3514.84.1.5.
18. Ehrlinger J, Johnson K, Banner M, Dunning D, Kruger J. 2008. Why the unskilled are unaware: further explorations of (absent) self-insight among the incompetent. Organ Behav Hum Decis Process 105:98–121. https://doi.org/10.1016/j.obhdp.2007.05.002.
19. Jensen PA, Moore R. 2008. Students' behaviors, grades and perceptions in an introductory biology course. Am Biol Teach 70:483–487. https://doi.org/10.2307/30163330.
20. Pazicni S, Bauer CF. 2014. Characterizing illusions of competence in introductory chemistry students. Chem Educ Res Pract 15:24–34. https://doi.org/10.1039/C3RP00106G.
21. Park YJ, Santos-Pinto L. 2010. Overconfidence in tournaments: evidence from the field. Theory Decis 69:143–166. https://doi.org/10.1007/s11238-010-9200-0.
22. Simons DJ. 2013. Unskilled and optimistic: overconfident predictions despite calibrated knowledge of relative skill. Psychon Bull Rev 20:601–607. https://doi.org/10.3758/s13423-013-0379-2.
23. Bol L, Hacker DJ, O'Shea P, Allen D. 2005. The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. J Exp Educ 73:269–290. https://doi.org/10.3200/JEXE.73.4.269-290.
24. Dembo MH, Jakubowski TG. 2003. The influence of self-protective perceptions on the accuracy of test prediction. AERA, Chicago, IL.
25. Helzer EG, Dunning D. 2012. Why and when peer prediction is superior to self-prediction: the weight given to future aspirations versus past achievement. J Pers Soc Psychol 103:38–53. https://doi.org/10.1037/a0028124.
26. Potgieter M, Ackermann M, Fletcher L. 2010. Inaccuracy of self-evaluation as additional variable for prediction of students at risk of failing first-year chemistry. Chem Educ Res Pract 11:17–24. https://doi.org/10.1039/C001042C.
27. Osterhage JL, Usher EL, Douin T, Bailey WM. 2019. Opportunities for self-evaluation increase student calibration in an introductory biology course. CBE Life Sci Educ 18:ar16. https://doi.org/10.1187/cbe.18-10-0202.
28. Bell P, Volckmann D. 2011. Knowledge surveys in general chemistry: confidence, overconfidence, and performance. J Chem Educ 88:1469–1476. https://doi.org/10.1021/ed100328c.
29. Dang NV, Chiang JC, Brown HM, McDonald KK. 2018. Curricular activities that promote metacognitive skills impact lower-performing students in an introductory biology course. J Microbiol Biol Educ 19:19.1.5. https://doi.org/10.1128/jmbe.v19i1.1324.
30. Siegesmund A. 2016. Increasing student metacognition and learning through classroom-based learning communities and self-assessment. J Microbiol Biol Educ 17:204–214. https://doi.org/10.1128/jmbe.v17i2.954.

31. Karpicke JD, Blunt JR. 2011. Retrieval practice produces more learning than elaborative studying with concept mapping. Science 331:772–775. https://doi.org/10.1126/science.1199327.
32. McDaniel MA, Anderson JL, Derbish MH, Morrisette N. 2007. Testing the testing effect in the classroom. Eur J Cogn Psychol 19:494–513. https://doi.org/10.1080/09541440701326154.
33. Roediger HL, Karpicke JD. 2006. Test-enhanced learning: taking memory test improves long-term retention. Psychol Sci 17:249–253. https://doi.org/10.1111/j.1467-9280.2006.01693.x.
34. Roediger HL, Agarwal PK, McDaniel MA, McDermott KB. 2011. Test-enhanced learning in the classroom: long-term improvements from quizzing. J Exp Psychol Appl 17:382–395. https://doi.org/10.1037/a0026252.
35. Adesope OO, Trevisan DA, Sundararajan N. 2017. Rethinking the use of tests: a meta-analysis of practice testing. Rev Educ Res 87:659–701. https://doi.org/10.3102/0034654316689306.
36. Rivers ML. 2020. Metacognition about practice testing: a review of learners' beliefs, monitoring, and control of test-enhanced learning. Educ Psychol Rev. https://doi.org/10.1007/s10648-020-09578-2.
37. Brame CJ, Biel R. 2015. Test-enhanced learning: the potential for testing to promote greater learning in undergraduate science courses. CBE Life Sci Educ 14:es4. https://doi.org/10.1187/cbe.14-11-0208.
38. Lopez EJ, Nandagopal K, Shavelson RJ, Szu E, Penn J. 2013. Self-regulated learning study strategies and academic performance in undergraduate organic chemistry: an investigation examining ethnically diverse students. J Res Sci Teach 50:660–676. https://doi.org/10.1002/tea.21095.
39. Tomanek D, Montplaisir L. 2004. Students' studying and approaches to learning in introductory biology. Cell Biol Educ 3:253–262. https://doi.org/10.1187/cbe.04-06-0041.
40. Sebesta AJ, Bray Speth E. 2017. How should I study for the exam? Self-regulated learning strategies and achievement in introductory biology. CBE Life Sci Educ 16:ar30. https://doi.org/10.1187/cbe.16-09-0269.
41. Zusho A, Pintrich PR, Coppola B. 2003. Skill and will: the role of motivation and cognition in the learning of college chemistry. Int J Sci Educ 25:1081–1094. https://doi.org/10.1080/0950069032000052207.
42. Brown PC, Roediger HL, McDaniel MA. 2014. Make it stick: the science of successful learning. Belknap Press, Cambridge, MA.
43. Hacker DJ, Bol L, Horgan D, Rakow E. 2000. Test prediction and performance in a classroom context. J Educ Psychol 92:160–170. https://doi.org/10.1037/0022-0663.92.1.160.
44. Kornell N, Rhodes MG. 2013. Feedback reduces the metacognitive benefit of tests. J Exp Psychol Appl 19:1–13. https://doi.org/10.1037/a0032147.
45. Cogliano M, Kardash CM, Bernacki ML. 2019. The effects of retrieval practice and prior topic knowledge on test performance and confidence judgments. Contemp Educ Psychol 56:117–129. https://doi.org/10.1016/j.cedpsych.2018.12.001.
46. Bol L, Hacker DJ. 2001. A comparison of the effects of practice tests and traditional review on performance and calibration. J Exp Educ 69:133–151. https://doi.org/10.1080/00220970109600653.
47. Koriat A, Sheffer L, Ma'ayan H. 2002. Comparing objective and subjective learning curves: judgments of learning exhibit increased underconfidence with practice. J Exp Psychol Gen 131:147–162. https://doi.org/10.1037/0096-3445.131.2.147.
48. Finn B, Metcalfe J. 2008. Judgments of learning are influenced by memory for past test. J Mem Lang 58:19–34. https://doi.org/10.1016/j.jml.2007.03.006.
49. Serra M, Dunlosky J. 2005. Does retrieval fluency contribute to the underconfidence-with-practice effect? J Exp Psychol Learn Mem Cogn 31:1258–1266. https://doi.org/10.1037/0278-7393.31.6.1258.
50. England BD, Serra MJ. 2012. The contributions of anchoring and past-test performance to the underconfidence-with-practice effect. Psychon Bull Rev 19:715–722. https://doi.org/10.3758/s13423-012-0237-7.
51. Lawson AE, Banks DL, Logvin M. 2007. Self-efficacy, reasoning ability, and achievement in college biology. J Res Sci Teach 44:706–724. https://doi.org/10.1002/tea.20172.
52. Blasiman RN, Dunlosky J, Rawson KA. 2017. The what, how much, and when of study strategies: comparing intended versus actual study behaviour. Memory 25:784–792. https://doi.org/10.1080/09658211.2016.1221974.