Opportunities for Self-Evaluation Increase Student Calibration in an Introductory Biology Course

Jennifer L. Osterhage,[†]* Ellen L. Usher,[‡] Trisha A. Douin,[‡] and William M. Bailey[§]
[†]Department of Biology, [‡]Department of Educational, School, and Counseling Psychology, and [§]Department of Physiology, University of Kentucky, Lexington, KY 40506

ABSTRACT

Accurate self-evaluation is critical for learning. Calibration describes the relationship between learners' perception of their performance and their actual performance on a task. Here, we describe two studies aimed at assessing and improving student calibration in a first-semester introductory biology course at a 4-year public institution. Study 1 investigated students' (*n* = 310) calibration (the difference between estimated and actual exam performance) across one semester. Students were significantly miscalibrated for the first exam: their predicted scores were, on average, significantly higher than their actual scores. The lowest-performing students had the most inaccurate estimates. Calibration improved with each exam. By the final exam, students underestimated their scores. We initiated a second study in the following semester to examine whether explicitly teaching students about self-evaluation strategies would improve their calibration and performance. Instruction in the experimental section (*n* = 290) focused on students' tendency to overestimate their abilities and provided retrieval-practice opportunities. Students in the experimental section showed better calibration and performance on the first exam compared with students in a control section taught by a different instructor during the same semester (*n* = 251). These findings suggest that simple instructional strategies can increase students' metacognitive awareness and improve their performance.

INTRODUCTION

Students' ability to distinguish what they know from what they do not yet know is critical for effective learning. However, students' perceived abilities are often misaligned with their actual knowledge (Serra and DeMarree, 2016). Almost all undergraduate students enter introductory courses expecting to earn a final grade of "A" or "B" (Beattie *et al.*, 2016). Students may become discouraged or draw incorrect conclusions (Stinebrickner and Stinebrickner, 2014) if their performance does not match their expectations. Metacognition, an ability to think about one's own thinking, is therefore crucial for academic success (Tanner, 2012). Metacognitive knowledge refers to what students know about learning, including their own learning processes, awareness of effective study strategies and when and why to use them, and ability to differentiate between knowing and not knowing (Schraw, 1998; Stanton *et al.*, 2015). The ability to evaluate one's understanding of information is an important component of metacognitive regulation (Lin and Zabrucky, 1998; Stone, 2000), whereby learners exercise control over their own learning by evaluating their own strengths and weaknesses, reflecting on their strategies, and planning better ways of doing things (Schraw, 1998; Zimmerman, 2002; Ambrose *et al.*, 2010). Although metacognitively competent students can identify gaps in their knowledge and adjust their strategies accordingly, metacognitively unaware students fail to realize the limits of their understanding and are at risk for self-regulatory and academic failure (Serra and Metcalfe, 2009). Given these significant implications for student learning, researchers have been increasingly interested in measuring and promoting students' metacognitive knowledge and regulatory skills (Tanner, 2012).

Kathryn E. Perez, *Monitoring Editor*
Submitted Oct 4, 2018; Revised Jan 7, 2019; Accepted Jan 17, 2019
CBE Life Sci Educ June 1, 2019 18:ar16
DOI:10.1187/cbe.18-10-0202

*Address correspondence to: Jennifer Osterhage (Jennifer.osterhage@uky.edu).

Calibration and the Dunning–Kruger Effect

Calibration describes the relationship between a judgment of one's performance and one's actual performance (Hattie, 2013; Schraw *et al*., 2013). Measures of student calibration have been used as a means to assess metacognitive knowledge (Hattie, 2009). Calibration has been measured in many ways, including by calculating the difference between students' predicted performance on a task and their actual performance (termed a discrepancy score). Well-calibrated students have low discrepancy scores, whereas poorly calibrated students have high discrepancy scores, and therefore do not accurately predict their performance. Calibration is essential to the metacognitive process and can influence academic success (Bembenutty, 2009). Calibrated students are more likely to earn higher course grades than students who are poorly calibrated (Garavalia and Gredler, 2002).

Evidence has also shown that the least competent individuals are the most likely to be overconfident in judgments of their performance. This cognitive bias, in which unskilled individuals rate their ability as higher than it is, was named the Dunning–Kruger effect (Kruger and Dunning, 1999). Kruger and Dunning found that the least competent individuals across multiple domains "grossly overestimated" their performance and ability. This observation is not new: Charles Darwin (1871) suggested more than a century ago that "ignorance more frequently begets confidence than does knowledge" (p. 3). Performance estimates seem to be based on long-standing self-views, which are only modestly related to prior performance (Ehrlinger and Dunning, 2003). Overconfidence is highest when students have inaccurate prior knowledge (van Loon *et al.*, 2013). The Dunning–Kruger effect has been observed in multiple studies across an array of skills and disciplines (Kruger and Dunning, 1999; Ehrlinger and Dunning, 2003; Ehrlinger *et al.*, 2008; Caputo and Dunning, 2005; Jensen and Moore, 2008; Pazicni and Bauer, 2014).

Research investigating student calibration and its effect on science classroom performance has shown that the lowest-performing students in introductory biology and chemistry courses were most likely to overestimate their specific content knowledge, exam performance, overall grades, and perceived class rank, while the highest-performing students tended to underestimate themselves in these areas (Jensen and Moore, 2008; Bell and Volckmann, 2011; Ziegler and Montplaisir, 2014; Siegesmund, 2016; Dang *et al.*, 2018). In an upper-level biology course, the lowest quartile of students perceived their knowledge to be similar to students in the upper quartile, despite a large gap in actual knowledge between the groups (Ziegler and Montplaisir, 2014). However, students' ability to evaluate their knowledge improved from a pretest to a posttest. In another study, the lowest-performing students increasingly overestimated their performance as the semester progressed (Pazicni and Bauer, 2014). Overconfidence has negative consequences on course grades. For example, one study demonstrated that the extent of overconfidence on a pretest predicted the likelihood of failing a first-semester chemistry course (Potgieter *et al.*, 2010). These data are especially important given the tendency for gateway undergraduate science courses to have high failing (D/F) or withdrawal (W) rates (Freeman *et al.*, 2011) and evidence that students may change their majors from a science to a non–science field because their grades in their science courses were substantially lower than they initially expected (Stinebrickner and Stinebrickner, 2014).

Self-Regulated Learning

Metacognitive regulation is a significant component of self-regulated learning (SRL; Schraw *et al.*, 2006). SRL is a process by which individuals develop goals, select learning strategies, and monitor their performance. Successful learners plan, organize, motivate themselves, self-monitor, and self-evaluate, modifying their tactics when necessary (Nilson, 2013). According to Zimmerman (2002), self-regulation involves a cyclical process with three key stages: forethought, performance, and self-reflection. The forethought phase includes planning and goal setting. The performance stage occurs during learning and includes self-control and self-observation. The self-reflection stage includes self-reaction and evaluative judgment. Metacognitively regulated students effectively assess a task, plan effective strategies to achieve their goals, continually monitor their understanding during the learning process, and make adjustments when necessary. Metacognitively unaware students may inaccurately assess a task, fail to make plans that match the task, ineffectively self-monitor during the learning process, and continue to apply ineffective strategies (Ambrose *et al.*, 2010).

One central component of being a self-regulated learner is the ability to evaluate one's own knowledge, a key component of metacognitive regulation. Students who lack these skills are often unaware of the limits of their knowledge. Therefore, the SRL cycle is disrupted when students are not aware of what they do and do not know (Ambrose *et al.*, 2010). When learners are overconfident, they may fail to realize when they should implement necessary self-regulatory strategies (Hadwin and Webster, 2013). They may make inappropriate decisions about how to study (Nelson and Narens, 1990) and may ignore valuable feedback and fail to take corrective actions (Hattie, 2013). The disruption of the SRL cycle due to miscalibration, therefore, has important implications for student achievement (DiBenedetto and Bembenutty, 2013; Dunlosky and Thiede, 2013).

The self-regulatory strategies used by students in undergraduate science courses have been described in some detail. Lopez *et al.* (2013) found that, in an undergraduate chemistry course, review-type strategies were common, but metacognitive strategies were not as widely used. Another study showed that self-regulatory strategies such as self-testing, monitoring understanding, and filling in gaps in understanding increased over time among students in an introductory chemistry course (Zusho *et al.*, 2003). Failure to use self-regulatory strategies has also been correlated with lack of success in introductory biology courses: students earning a "D" or "F" on the first exam in an introductory biology course reported lower usage of effective SRL strategies such as self-evaluation, planning, and seeking assistance (Sebesta and Speth, 2017). These studies underscore the importance of the development of SRL skills for success in college science courses.

Improving Self-Judgment Accuracy through SRL and Metacognitive Instruction

Specific study strategies have been shown to improve metacognitive competence and exam performance (Siegesmund, 2016). Despite the importance of metacognition on knowledge gains, undergraduate students are largely unaware that specific study strategies are associated with increased metacognitive awareness, whereas others are much less effective (Brown *et al.*, 2014;

Sebesta and Speth, 2017). For example, Karpicke and Blunt (2011) showed that repeated review of material gives students the illusion that they know material better than they do, enhancing miscalibration. When students reread notes, their familiarity with the information gives them a false sense that they fully understand the underlying concepts (Brown *et al.*, 2014). In contrast, retrieval practice—actively recalling information from memory—reduces overconfidence and improves performance on subsequent assessments. Retrieval practice has at least two benefits. First, it may help students recognize gaps in their knowledge that they can address through additional studying. Second, retrieval activities cause the brain to consolidate what is learned, thereby strengthening the connections between what is being learned and prior knowledge (Brown *et al.*, 2014). Recall tests, therefore, can enhance SRL and memory and decrease miscalibration. Strategies such as classroom-based learning communities, which provide frequent feedback and address prior misconceptions, have also been shown to help students improve the accuracy of their self-judgments (Schraw *et al.*, 2006; Hattie, 2013; Siegesmund, 2016).

Explicit instruction in SRL, metacognition, and effective study strategies can improve metacognitive skills and calibration (Winne and Hadwin, 1998; McCabe, 2011; Zimmerman *et al.*, 2011). Instructor-led activities can promote general metacognitive skill development or focus on specific aspects of the SRL cycle (for a review, see Ambrose *et al.*, 2010). For example, Nietfeld *et al.* (2006) showed that students who received feedback on their calibration and self-monitoring activities had increased calibration and class performance in an undergraduate educational psychology course. Instruction in the use of "enhanced answer keys," which included explanations of correct answers, details about how questions were scored, and additional reflection questions designed to engage students in metacognition, resulted in significantly higher learning gains for students enrolled in an introductory biology course (Sabel *et al.*, 2017). Curricular activities designed to promote metacognitive skill development in an introductory biology course were associated with an increase in the accuracy of postdiction (i.e., after exam) estimates of exam scores as the semester progressed (Dang *et al.*, 2018).These studies indicate that interventions aimed at increasing the accuracy of student judgment can positively affect performance on summative assessments.

Students' ability to gauge their preparedness for summative assessments is of particular importance due to the repercussions that inaccurate judgments have on learning, self-regulation, and final course grade. However, relatively few studies have assessed how students' calibration accuracy as measured by the difference in predicted and actual exam scores might change over time during their undergraduate science courses. Moreover, few studies have examined whether explicit instruction on metacognitive awareness might be related to students' calibration in a large undergraduate introductory science course. To explore this further, in study 1, we investigated patterns in biology students' metacognitive awareness (i.e., exam calibration). In study 2, we then examined the efficacy of an instructional intervention designed to improve metacognitive awareness. We asked the following research questions:

1. How does student calibration change over the semester in an introductory biology course? (study 1)
2. Do introductory biology students exhibit the Dunning–Kruger effect (i.e., are the lowest-performing students the most likely to be overconfident)? (study 1 and study 2)
3. Can an instructional intervention improve students' calibration early in the semester? (study 2)

METHODS

Participants

Both studies were conducted in large-enrollment introductory biology courses at a 4-year public institution in the southeastern United States. Participants in study 1 (n = 290) were enrolled in one course section during the Fall semester. Participants in study 2 (n = 541) were enrolled in either a control section (n = 251) or an experimental section (n = 290) in the Spring semester of the same academic year. To maintain consistency across semesters and sections, we excluded from the analysis the data from students who did not complete the course (i.e., withdrew before the end of the semester).

Course Description and Setting

Introductory Biology I is a required course for the biology major and many other science and pre–health majors across the university. Up to four sections of the course are taught each semester. 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. Some activities to promote self-evaluation are routinely embedded throughout the course: 1) clicker questions for which students can see the percentage of classmates who chose each answer, 2) group quizzing in which students are encouraged to think about how many of their answers they changed after group discussion, and 3) practice questions and a practice exam with answer keys that include feedback about correct and incorrect answers. Effective study strategies are discussed in class, and documents describing effective study habits are available on the learning management system. Data collection was approved through the university's institutional review board (14-0959-P4S).

Study 1

Design and Procedures. Students in study 1 were visited by researchers J.L.O. and T.A.D. at the beginning of each of four exam periods, which were equally spaced across the semester. Exams were administered to all students enrolled in the course on the same day during the same common hour exam time. Researchers distributed a half-page questionnaire and asked students to fill it out before completing their exam (see the Supplemental Material). Questionnaires were collected separately from exams at the end of the exam period. Students wrote their student ID on the questionnaires.

Data Sources. Data included predicted exam scores, actual exam scores, and their discrepancies.

Exam Score Predictions. On each pre-exam questionnaire, students were asked to estimate their exam scores (i.e., the percentage correct they thought they would earn on the exam).

Biology Exam Scores. Instructors provided a list of exam scores and student ID numbers to researchers (T.A.D.) unaffiliated

with the course. The first three biology exams consisted of 50 multiple-choice questions worth 2 points each. The final exam consisted of 100 multiple-choice questions (50 from new material and 50 cumulative) and was worth 150 points. Final exam scores were converted to percentages, thus placing each exam on the same scoring metric.

Calibration Scores. Discrepancy scores were calculated as the difference between students' predicted and actual 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. The absolute value of the discrepancy score was used to provide an indication of how "off" a student's estimate was, without indicating the direction of the miscalibration (i.e., over- or underestimation). Students whose absolute-value discrepancy score was 10 or greater (i.e., one letter grade) were considered to be "miscalibrated."

Analysis. To answer our first research question (RQ1) related to changes in students' discrepancy scores across the semester, we used repeated-measures one-way analysis of variance (ANOVA) with Tukey's multiple-comparisons posttest. For each exam, students' predicted and actual scores were plotted against their actual performance percentile to examine whether the lowest-performing students were also the most likely to be overconfident (i.e., RQ2). The best-fit lines of the data were also plotted for each exam. The correlation coefficient (R^2) value between actual and predicted scores was calculated. The strength of the correlation between the best-fit actual and predicted performance lines was compared using a Fisher *r*-to-*z* transformation. Significant outliers were identified by the iterate Grubbs' method and excluded from analysis.

Study 2

Design and Procedures. Study 2 was designed to test whether implementing an instructional intervention before the first introductory biology exam might improve students' calibration. As noted earlier, this study involved data collected from students at the same institution enrolled in two sections of introductory biology during the Spring semester. The sections followed the same course format as in study 1. The same exam was given to students in both sections of study 2. A second practice exam was distributed to all students in study 2. Because the instructor implementing the instructional intervention was the same instructor whose students participated in study 1, we used the data collected in study 1 for statistical comparison in study 2.

Control Section. Students not receiving the instructional intervention were enrolled in one course section of the Spring semester (n = 251). This course section was taught by an instructor with similar experience to the instructor in the experimental section. All introductory biology instructors share materials and exams. Therefore, the same instructional techniques described for study 1 (i.e., clicker questions, practice questions, and group quizzes) were used by the instructor in this control section.

Experimental Section. The experimental section (n = 290) was taught by the same instructor who taught the participating course section in study 1. Two instructional intervention activities, described in the following sections, were offered in the experimental section but not in the control section. Otherwise, the same instructional techniques described for the control section were implemented in the experimental section. The instructional interventions, which took place before the first exam, were designed to increase both students' metacognitive knowledge and regulation by acquainting students with their habitual strategies (i.e., poor calibration and marginally effective studying skills), presenting them with counterevidence from two studies from the learning sciences literature, and providing them with opportunities to practice new strategies.

Activity 1. The goal of this activity was to illustrate the tendency of individuals to overestimate abilities and discuss self-regulatory learning strategies. On the first day of class, students were given a 25-question pretest over concepts to be covered throughout the semester. The next class day, students were asked (via clicker) to estimate their performance on the pretest that was administered on the first day of class. Fifty-seven percent of the class estimated that they earned above a 70% and only 8% of the class estimated that they earned below 50%. Students were then shown the actual distribution of scores. Only 11% of the students actually earned above a 70%, and 32% of students earned less than 50%. This activity was used to launch a discussion (less than 20 minutes) of the general phenomenon that individuals tend to overestimate their abilities. Students were also shown the mean discrepancy score on the first exam from study 1 (see *Results*). The instructor then outlined self-regulatory learning strategies that have been shown to promote more accurate self-judgment (Winne and Hadwin, 1998; McCabe, 2011). See the Supplemental Material for slides used to facilitate discussion.

Activity 2. The goal of this activity was to provide additional opportunities for retrieval practice and to discuss how to use feedback effectively. During the second week of class, students were polled via clicker about the study strategy they planned to use the most in the course (of a list of five strategies provided; see the Supplemental Material). Reviewing notes was the most popular answer. This prompted a short discussion (less than 20 minutes) about a study by Karpicke and Blunt (2011), who found that reviewing notes gives students the illusion that they understand material better than they do and that retrieval-practice activities improve performance on subsequent assessments. The instructor shared findings showing that learning is enhanced when study time is dedicated to elaborative activities (e.g., making models) and practice. Students were encouraged to complete both practice exams and were advised to look at the answer keys only after they attempted the entire practice exam on their own to objectively assess their current understanding. The instructor discussed how feedback from the practice exams could be used to adjust study strategies. As in the previous semester, answer keys included feedback about correct and incorrect answer choices. See the Supplemental Material for slides used to facilitate discussion.

Study 1 Comparison Section. To ensure that any differences observed in study 2 were due to the instructional intervention and not to other factors, we compared our findings with scores from participants in study 1, which involved students enrolled

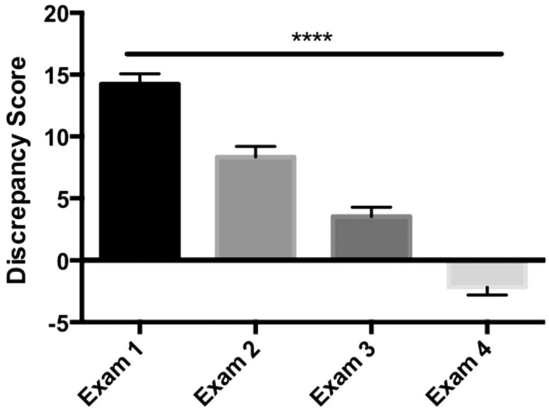

FIGURE 1. Student calibration improves over time in study 1. Mean difference ± SEM between score predicted before exam and actual score for each exam is plotted.

in the course the previous semester who had the same instructor as the instructor who implemented the instructional intervention in study 2. Unlike the experimental section participants in study 2, participants in study 1 did not receive the instructional intervention despite having had the same instructor. Using data from these students allowed us to control for differences by instructor and student demographics between the Fall and Spring semesters.

Data Sources. Data in study 2 were collected in the same manner as in study 1 (see the Supplemental Material); however, because discrepancy scores in study 1 were highest for the first exam, study 2 focused only on the first exam. Data were gathered for only the first exam of the semester in study 2. The only data used from study 1 were from exam 1. Predicted exam scores and actual exam scores were gathered, and the discrepancy between the two was calculated.

Analysis

Our third research question aimed to evaluate whether an instructional intervention could improve students' calibration early in the semester. We used one-way ANOVA with post hoc Tukey's *t* test comparisons to compare students' exam 1 discrepancy scores across conditions (experimental, control section, study 1 comparison section). As in study 1, we examined graphs of students' predicted and actual scores plotted against actual performance percentile to examine whether the lowest-performing students were also the most likely to be overconfident (i.e., RQ2).

RESULTS

Study 1

Does Student Calibration Change over Time? We investigated students' ability to predict their scores on exams by calculating the difference between students' earned exam scores and the scores predicted before taking the exam; this difference was called a discrepancy score. The mean discrepancy score was 14 points for the first exam (M_{actual} = 71%; $M_{estimated}$ = 85%), indicating that, on average, students overestimated their performance on the first exam. However, the mean discrepancy score steadily decreased as the semester progressed (Figure 1). A repeated-measures ANOVA test revealed that the difference between discrepancy scores across each subsequent exam (i.e., from exam 1 to exam 2, etc.) was statistically significant ($p < 0.05$) for all exam pairs. By the final exam, students as a whole underestimated their performance by an average of 2.2%, indicating that they became more calibrated over time.

Is Miscalibration Related to Performance? To initially examine patterns of miscalibration at the individual student level, we calculated the percentage of miscalibrated students (i.e., those with absolute-value discrepancy scores ≥10) for each exam. Notably, the percentage of students who underestimated their score increased with each exam, while the percentage of students who overestimated their performance decreased with each exam (see Table 1).

Actual and predicted scores graphed against percentile rank of actual performance, along with the best-fit lines of the data, were plotted for each exam. The Dunning–Kruger effect was observed in this introductory biology learning context: the lowest-performing students were least calibrated (most discrepant) when predicting their scores on the first exam (Figure 2, left). Students performing in the lowest quartile overestimated their scores on exam 1 by an average of 32 points.

Correlation coefficients between predicted and actual exam scores were significantly higher at exam 3 compared with exam 1. By the final exam, many of the lowest-performing students had improved their exam score predications (Figure 2, right). The lowest-performing quartile of students overestimated their performance by an average of 6 points on the final exam. By the final exam, the correlation between predicted and actual scores, although still significantly stronger than it was at exam 1, plateaued due to the increased proportion of students who underestimated their scores. These findings indicate that, as the semester progressed, the lowest-performing students both adjusted their predictions (Figure 2, compare slopes of predicted score lines between panels) and improved their exam performance (Figure 2, compare slopes of actual score lines between panels).

A plot of the raw discrepancy scores for each student revealed that most students trended toward increased calibration as the semester progressed (Figure 3). Five percent of students, however, continued to be miscalibrated at every time point (red lines). These students were among the lowest performing in the class.

TABLE 1. The percentage of miscalibrated students changes over time (study 1)

	Exam 1	Exam 2	Exam 3	Exam 4
% Miscalibrated	59.7	41.7	38.4	27.5
% Overestimated performance	59.7	35.5	26.5	8.1
% Underestimated performance	0.0	6.2	11.8	19.4

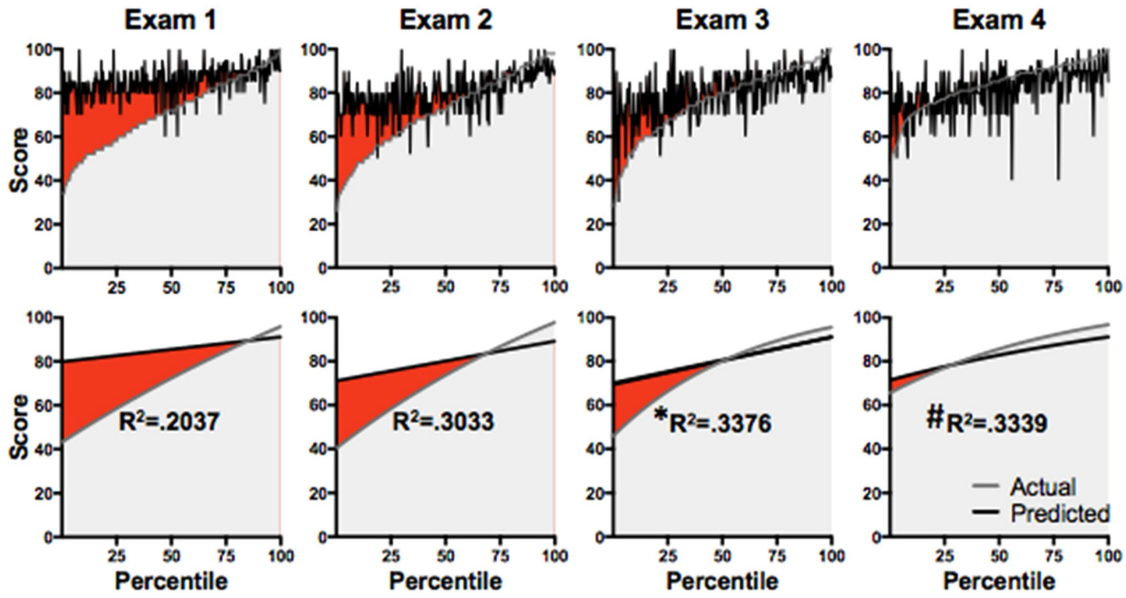

FIGURE 2. Correlation between predicted and actual scores improves with time in study 1. Predicted and actual scores graphed by percentile rank of actual scores for exams 1–4 are plotted (top). Trends are visualized by best curve fit lines (bottom). The area between curves (red) represents overestimation. The correlation coefficient (R^2) value between actual and predicted scores was calculated. Exam 3 and exam 4 both had significantly stronger correlations compared with exam 1 when outliers were excluded. The Fisher *r*-to-*z* transformation was used to identify significant difference in strength of correlation (*, $p \leq 0.05$; #, $p \leq 0.05$) when four outliers who underestimated their score by an average of 40 points were excluded.

Study 2

Can an Instructional Intervention Improve Students' Calibration? Because discrepancy scores in study 1 were highest for the first exam of the semester, study 2 focused on whether an instructional intervention might improve students' calibration on the first exam. We first compared the exam 1 discrepancy scores of students in the two sections of the second semester (who all completed the same exam). Findings indicated that students who were in the instructional intervention section (i.e., experimental group) had a significantly lower average discrepancy score (i.e., were better calibrated) on the first exam than did students in the control section (Figure 4). The percentage of students in the experimental section who overestimated their performance was significantly lower than in the control section (see Table 2). These findings suggest that the intervention activities helped students become better estimators of their own performance.

To consider additional empirical support for this inference, we first compared the scores of students in the control section with those of students in study 1 (i.e., who had the same instructor as the study 2 experimental group but had not received the intervention). The mean exam 1 discrepancy score was not statistically different between these two groups, indicating similar levels of miscalibration in sections without instructional intervention. We next examined whether students taught by the same instructor in study 1 (receiving no instructional intervention) and study 2 (receiving instructional intervention) differed in their calibration levels. Indeed, the instructor's students were significantly better calibrated (i.e., less discrepant) in their estimates in study 2 ($M = 3.5\%$) than in study 1 ($M = 13.5\%$; Figure 4).

Is Miscalibration Related to Performance? As in study 1, we found evidence in study 2 of the Dunning–Kruger effect. Students in the lowest-performing quartile of each section were the most miscalibrated (Figure 5, right). However, the magnitude of miscalibration was significantly smaller in the experimental group than in the other two groups. The lowest-performing quartile of students in the experimental group overestimated their scores on the first exam by 18 points, a 14-point improvement over the same instructor's students in study 1. The highest-performing quartile of students in study 2 underestimated their scores by 6 points, whereas students in the second and third quartiles were well calibrated for the first exam.

The improved calibration cannot be explained by a difference in the slope of the predicted lines (compare predicted score lines in Figure 5, bottom). The difference in calibration, therefore, is mainly due to improved performance on the first exam by students who received the instructional intervention. Their average scores on exam 1 were 7 points higher than the scores earned by

TABLE 2. Early intervention affects calibration and early exam performance

	Study 1	Study 2: control section	Study 2: experimental section
% Miscalibrated	59.7	58.7	36.6
% Overestimated performance	59.7	53.9	28.0
% Underestimated performance	0.0	4.8	8.6

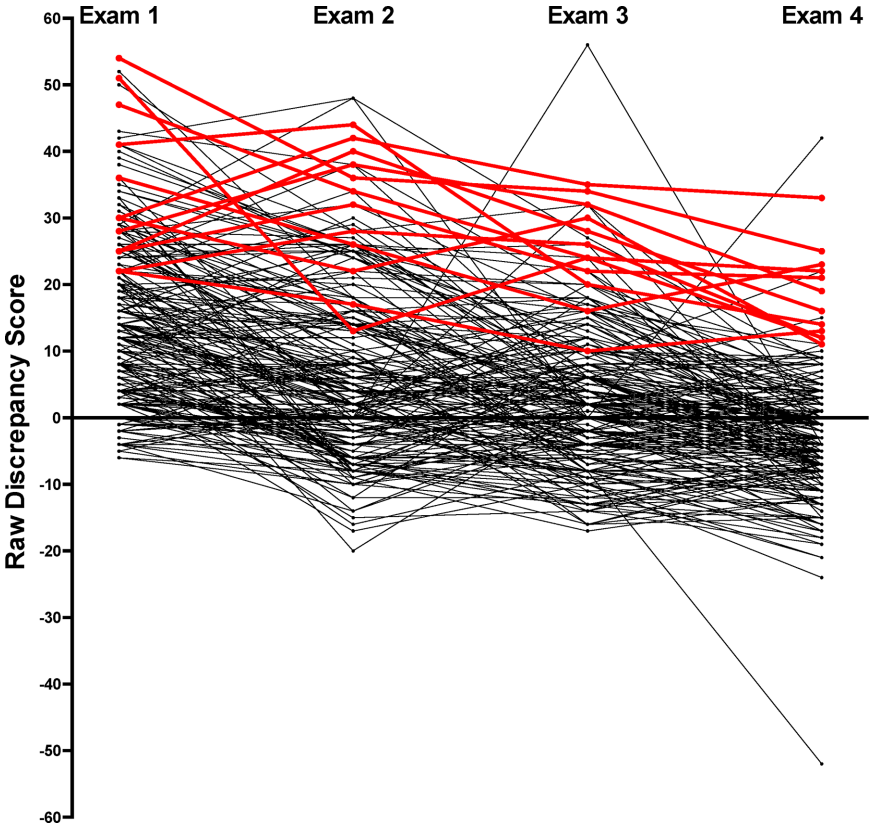

FIGURE 3. Calibration over time varies between students in study 1. Raw discrepancy scores for each student are plotted for each exam. Red lines are students who overestimated their performance by at least 10 points for each exam.

students in the previous semester and 8 points higher than scores of students in the control section. Similarly, compared with the control section, fewer students in the experimental section scored below 50% on the first exam, suggesting that the instructional intervention activities were effective for even the lowest-performing students (Table 3).

Taken together, these data suggest that instructor interventions to foster self-evaluation are associated with increased calibration and exam performance early in the semester.

DISCUSSION

Accurate self-judgment is a hallmark of the metacognitive, self-regulated learner. Miscalibration disrupts the SRL cycle and has been shown to negatively affect student achievement in undergraduate science courses. In the first study, we investigated how student calibration changed over the course of one semester in an introductory biology course. Although most students began the course greatly overestimating their performance, the accuracy of most students' exam predictions improved over the course of the semester. By the final exam, students had a tendency to underestimate their performance on average. We speculate that familiarity with the course and exam structure, feedback from previous exam performance, and/or self-reflective activities built into the course (practice questions, in-class clicker questions, and group quizzing) contributed to improved calibration over time. These findings are in agreement with those reported in a study in which the accuracy of students' postdiction performance estimates similarly improved as the semester progressed (Dang *et al*., 2018). Although postdiction estimates of exam performance provide important information about metacognitive awareness, this measure does not capture student perceptions of their preparedness for summative assessments before they occur. Predictions of performance are important indicators of students' perceived level of preparedness for exams. Overconfident students typically have studied ineffectively and failed to implement necessary self-regulatory strategies, which can have negative consequences on course grades and achievement.

Several factors might explain why students improve their calibration over time. Specifically, more accurate self-judgment, an increase in actual performance, or a combination of the two would reduce miscalibration. Metacognitively aware students should be able to gauge their understanding of the material accurately, resulting in low discrepancy scores regardless of their earned scores. Discrepancy scores can also decrease with increased exam performance, even without adjustments in predicted scores, due to a restriction of range in possible discrepancy. When predicted scores are plotted relative to actual performance percentile, a relatively flat line would result if most students make similar predictions about their scores. However, metacognitive awareness should result in a sloped line; that is, the lowest-performing students should naturally predict lower scores and the higher-performing students should predict higher scores. In study 1, the slope of the best-fit "predicted" line was greater at the final exam than at the first exam (Figure 2), indicating that students' calibration changes over time in part because their self-evaluation becomes more accurate. In addition, fewer students earned low exam scores as the semester progressed, indicating that improved performance was also a factor in the increased association between predicted and earned scores over time. This finding indicates that improvements in calibration accuracy are positively associated with exam performance. Even so, our findings indicated that, as found in previous studies, the lowest-performing students had the most inaccurate performance judgments (i.e., the Dunning–Kruger effect; Kruger and Dunning, 1999). Moreover, some of the lowest-performing students who completed the course continued to be miscalibrated as the semester progressed. It will be important to determine the causes and remedies of persistent miscalibration. If risk factors for failure to complete the course and extreme or persistent miscalibration could be identified early in the semester, targeted intervention might be possible.

In a follow-up investigation, study 2, we examined whether an intervention aimed to promote more accurate self-evaluation could help introductory biology students become better calibrated in their judgments. We used two instructional strategies

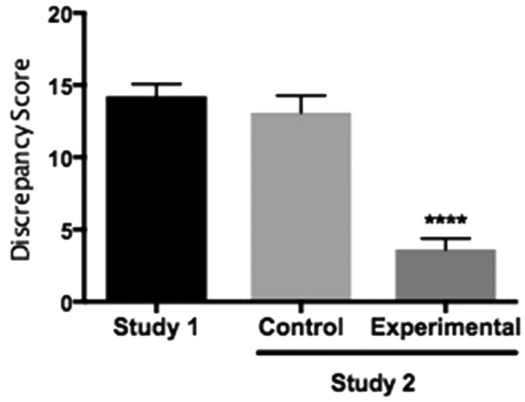

FIGURE 4. Interventions to foster self-evaluation are associated with increased calibration on exam 1 in study 2. Mean difference ± SEM between exam 1 discrepancy scores for three introductory biology sections. Means for study 2 experimental section remained significantly different from those for study 1 and study 2 control section. ****, $p < 0.001$

designed to help students become better calibrators. These strategies were designed to enhance both metacognitive knowledge and metacognitive regulation. Our findings suggest that either or both strategies were effective at helping students make more accurate predictions of their performance on the first biology exam. The first strategy involved increasing students' awareness of the tendency to overestimate their abilities, one element of metacognitive knowledge. The second strategy communicated to students the importance of retrieval-practice activities for their learning. Retrieval practice can help students in monitoring and evaluating their learning, critical components of metacognitive regulation. Our findings revealed that the difference in calibration after the instructor-led intervention was mainly due to increased performance on the first exam, not a difference in predicted scores between sections. This likely indicates that the instructor interventions helped improve students' learning, resulting in a closer match between their performance estimates and actual scores. These results are in alignment with those from previous findings that prompting students to use metacognitive approaches can change study habits and, in turn, learning (Stanton *et al.*, 2015). This has particularly important implications given recent studies that show that performance early in the semester is associated with subject matter self-efficacy and second-semester retention (Wright *et al.*, 2013; Ainscough *et al.*, 2016).

Another aspect of the instructional intervention involved providing students with performance feedback early in the course that addressed their tendency to inflate performance estimates. Not surprisingly, our findings support previous studies that have shown that valid feedback is essential when making self-judgments. Instructor-driven interventions early in the semester are particularly important, because they can influence self-judgment before summative assessments occur. We were encouraged that the simple strategies described here were correlated with an improvement in calibration and performance on the first exam. The feedback given to students in the instructional intervention were instructor-led, but this need not be the case.

Limitations and Future Directions

Although the results presented here were encouraging, the studies were limited in several ways. First, the fact that we combined

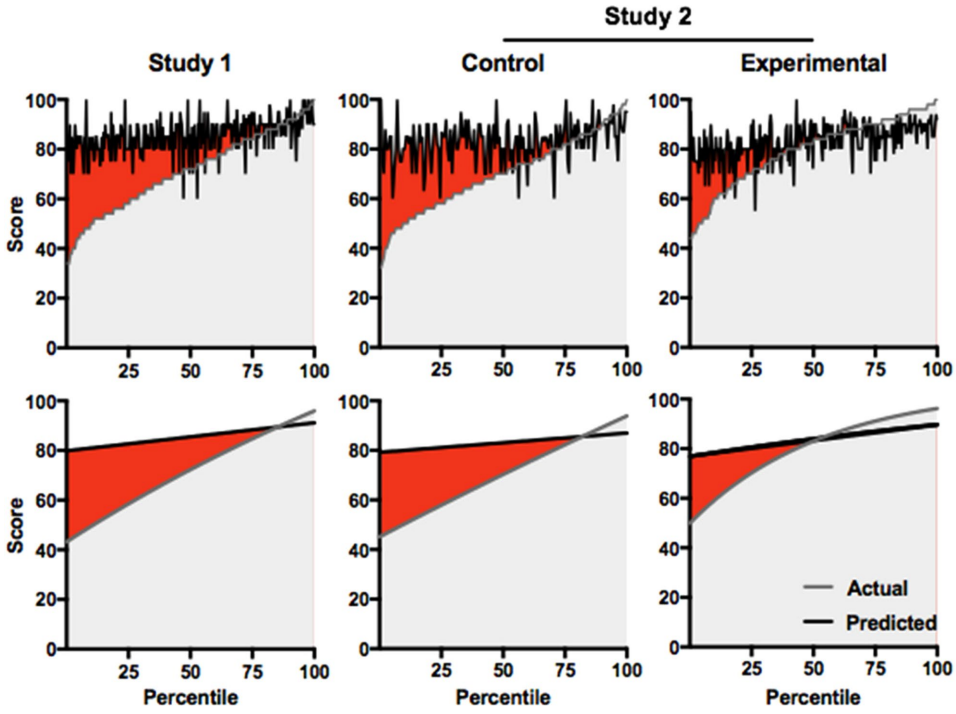

FIGURE 5. Interventions to foster self-evaluation are correlated with better calibration on the first exam in study 2. Predicted and actual scores graphed as a function of percentile rank of actual scores for exam 1 are plotted (top). Trends are visualized by best curve fit lines (bottom). The area between curves (red) represents overestimation.

TABLE 3. Instructor-led interventions are correlated with higher scores on exam 1

	Study 1	Study 2: control section	Study 2: intervention section
Exam average (%)	71	70	78
Percentage of students earning below 50%	9.5	8.5	3.7

interventions in study 2 makes it difficult to know the causal mechanisms through which students' calibration improved. It would be helpful to assess the magnitude of effect for any one part of the intervention (e.g., direct emphasis on the importance of calibration vs. practice tests). In addition, we cannot rule out the possibility that demographic differences between sections of the course and between semesters may have affected our results. It is also possible that "wishful thinking" contributed to early-semester miscalibration as opposed to, or in combination with, the Dunning–Kruger effect (Serra and DeMarree, 2016).

We did not examine students' use of study strategies over the course of the semester in either study. It would be interesting to know whether strategy use changed as the semester progressed and whether specific strategies were associated with increased calibration. Comparing strategies used by students in each section would also point to the specific approaches that improve calibration. To make comparisons between sections, these studies tracked only those students who completed each exam in the semester, thereby excluding those students who withdrew from the course before the end of the semester. The withdrawal rate in study 1 was 6.4%. In study 2, 4.92% of students withdrew from the experimental section, while 13.6% of students withdrew from the control section. Future studies could explore the calibration and study strategy usage of students who withdrew from the course. Similarly, we limited our investigation of calibration to the first exam of study 2. Investigating the longer-term effects of the instructional intervention would be useful.

Implications for Instructors

The findings of this investigation suggest several implications for science instructors. Explicit metacognitive instruction, especially when embedded in a course and adapted to specific learning contexts, can improve student performance (Zohar and David, 2009). Strategies to increase metacognitive awareness do not need to be time-intensive. For example, this study showed that the simple act of making students aware of their tendency to overestimate their abilities may help them make more accurate self-judgments before the first formative course assessment. Stressing the importance of retrieval-practice activities and the provision of ample opportunities for practice and retrieval (e.g., questions during class, practice questions, and practice exams) can both help students realize what they have and have not learned and help them consolidate and connect course concepts (Brown *et al.*, 2014). Practice exams seem particularly important, in that they give students a measure of their current understanding and allow students to learn from their mistakes without a negative effect on their course grade. Instructors can encourage students to use feedback from practice exams to focus on weak points and adjust their regulatory strategies as necessary. Communication of effective metacognitive and SRL strategies such as those described here make learning more accessible to a larger number of students (Pintrich, 2002). These efforts have important implications for student retention and success in science, technology, engineering, and mathematics fields.

ACKNOWLEDGMENTS

We thank the facilitators of the American Society for Microbiology's Biology Scholars program, especially Marcy Kelly, for important insights and feedback. We also thank the students who participated in this study. This work was supported by an HHMI Sustaining Excellence-2014 grant (#52008116).

REFERENCES

Ainscough, L., Foulis, E., Colthorpe, K., Zimbardi, K., Robertson-Dean, M., Chunduri, P., & Lluka, L. (2016). Changes in biology self-efficacy during a first-year university course. *CBE—Life Sciences Education*, *15*(2), ar19. doi: 10.1187/cbe.15-04-0092

Ambrose, S. A., Bridges, M. W., DiPietro, M., Lovett, M. C., & Norman, M. K. (2010). *How learning works: Seven research-based principles for smart teaching*. San Francisco: Jossey-Bass.

Beattie, G., Laliberté, J. W. P., & Oreopoulos, P. (2016). Thrivers and divers: Using non-academic measures to predict college success and failure. *Economics of Education Review*, *62*, 170–182. doi: 10.1016/j.econedurev.2017.09.008

Bell, P., & Volckmann, D. (2011). Knowledge surveys in general chemistry: Confidence, overconfidence, and performance. *Journal of Chemical Education*, *88*(11), 1469–1476. doi: 10.1021/ed100328c

Bembenutty, H. (2009). Three essential components of college teaching: Achievement calibration, self-efficacy, and self-regulation. *College Student Journal*, *43*(2), 562–570.

Brown, P. C., Roediger, H. L., & McDaniel, M. A. (2014). *Make it stick: The science of successful learning*. Cambridge, MA: Belknap Press. doi: 10.419/9780674419377-008

Caputo, D., & Dunning, D. (2005). What you don't know: The role played by errors of omission in imperfect self-assessments. *Journal of Experimental Social Psychology*, *41*(5), 488–505. doi: 10.1016/j.jesp.2004.09.006

Dang, N. V., Chiang, J. C., Brown, H. M., & McDonald, K. K. (2018). Curricular activities that promote metacognitive skills impact lower-performing students in an introductory biology course. *Journal of Microbiology and Biology Education*, *19*, 1–9. doi: 10.1128/jmbe.v19i1.1324

Darwin, C. (1871). *The descent of man, and selection in relationship to sex*. London: John Murray.

DiBenedetto, M. K., & Bembenutty, H. (2013). Within the science pipeline: Self-regulated learning and academic achievement among college students in science classes. *Learning and Individual Differences*, *2*, 218–224.

Dunlosky, J., & Thiede, K. W. (2013). Four cornerstones of calibration research: Why understanding students' judgments can improve their achievement. *Learning and Instruction*, *24*, 58–61. doi: 10.1016/j.learninstruc.2012.05.002

Ehrlinger, J., & Dunning, D. (2003). How chronic self-views influence (and potentially mislead) estimates of performance. *Journal of Personality and Social Psychology*, *84*, 5–17. doi: 10.1037/0022-3514.84.1.5

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. *Organizational Behavior and Human Decision Processes*, *105*, 98–121. doi: 10.1016/j.obhdp.2007.05.002

Freeman, S., Haak, D., & Wenderoth, M. P. (2011). Increased course structure improves performance in introductory biology. *CBE—Life Sciences Education*, *10*(2), 175–186. doi: 10.1187/cbe.10-08-0105

Garavalia, L. S., & Gredler, M. E. (2002). An exploratory study of academic goal setting, achievement calibration and self-regulated learning. *Journal of Instructional Psychology*, *29*(4), 221–230.

Hadwin, A., & Webster, E. (2013). Calibration in goal setting: Examining the nature of judgments of confidence. *Learning and Instruction*, *24*, 37–47. doi: 10.1016/j.learninstruc.2012.10.001

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses related to achievement*. New York: Routledge.

Hattie, J. (2013). Calibration and confidence: Where to next? *Learning and Instruction*, *24*, 62–66. doi: 10.1016/j.learninstruc.2012.05.009

Jensen, P. A., & Moore, R. (2008). Students' behaviors, grades and perceptions in an introductory biology course. *American Biology Teacher*, *70*(8), 483–487. doi: 10.2307/30163330

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*(6018), 772–775. doi: 10.1126/science.1199327

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121–1134. doi: 10.1037/0022-3514.77.6.1121

Lin, L. M., & Zabrucky, K. M. (1998). Calibration of comprehension: Research and implications for education and instruction. *Contemporary Educational Psychology*, *23*(4), 345–391. doi: 10.1006/ceps.1998.0972

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

McCabe, J. (2011). Metacognitive awareness of learning strategies in undergraduates. *Memory & Cognition*, *39*(3), 462–476. doi: 10.3758/s13421-010-0035-2

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In Bower, G. H. (Ed.), *The psychology of learning and motivation* (pp. 125–173). New York: Academic.

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2006). The effect of distributed monitoring exercises and feedback on performance, monitoring accuracy, and self-efficacy. *Metacognition and Learning*, *1*, 159–179. doi: 10.1007/s10409-006-9595-6

Nilson, L. (2013). *Creating self-regulated learners: Strategies to strengthen students' self-awareness and learning skills*. Sterling, VA: Stylus.

Pazicni, S., & Bauer, C. F. (2014). Characterizing illusions of competence in introductory chemistry students. *Chemistry Education Research and Practice*, *15*, 24–34. doi: 10.1039/C3RP00106G

Pintrich, P. R. (2002). The role of metacognitive knowledge in learning, teaching, and assessing. *Theory into Practice*, *41*(4), 219–225. doi: 10.1207/s15430421tip4104_3

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. *Chemistry Education Research and Practice*, *1*, 17–24. doi: 10.1039/C001042C

Sabel, J. L., Dauer, J. T., & Forbes, C. T. (2017). Introductory biology students' use of enhanced answer keys and reflection questions to engage in metacognition and enhance understanding. *CBE—Life Sciences Education*, *16*(3), ar40. doi: 10.1187/cbe.16-10-0298

Schraw, G. (1998). Promoting general metacognitive awareness. *Instructional Science*, *26*, 113–125.

Schraw, G., Crippen, K. J., & Hartley, K. (2006). Promoting self-regulation in science education: Metacognition as part of a broader perspective on learning. *Research in Science Education*, *36*, 111–139. doi: 10.1007/s11165-005-3917-8

Schraw, G., Kuch, F., & Gutierrez, A. (2013). Measure for measure: Calibrating ten commonly used calibration scores. *Learning and Instruction*, *24*, 48–57. doi: 10.1016/j.learninstruc.2012.08.007

Sebesta, A. J., & Speth, E. B. (2017). How should I study for the exam? Self-regulated learning strategies and achievement in introductory biology. *CBE—Life Sciences Education*, *16*(2), ar30. doi: 10.1187/cbe.16-09-0269

Serra, M. J., & DeMarree, K.G. (2016). Unskilled and unaware in the classroom: College students' desired grades predict their biased grade predictions. *Memory & Cognition*, *44*(7), 1127–1137. doi: 10.3758/s13421-016-0624-9

Serra, M. J., & Metcalfe, J. (2009). Effective implementation of metacognition. In Hacker, D. J., Dunlosky, J., & Graesser, A. C. (Eds.), *Handbook of metacognition in education* (pp. 278–298). New York: Routledge.

Siegesmund, A. (2016). Increasing student metacognition and learning through classroom-based learning communities and self-assessment. *Journal of Microbiology & Biology Education*, *17*(2), 204–214. doi: 10.1128/jmbe.v17i2.954

Stanton, J. D., Neider, X. N., Gallegos, I. J., & Clark, N. C. (2015). Differences in metacognitive regulation in introductory biology students: When prompts are not enough. *CBE—Life Sciences Education*, *14*(2), ar15. doi: 10.1187/cbe.14-08-0135

Stinebrickner, R., & Stinebrickner, T. R. (2014). A major in science? Initial beliefs and final outcomes for college major and dropout. *Review of Economic Studies*, *81*, 426–472. doi: 10.1093/restud/rdt025

Stone, N. J. (2000). Exploring the relationship between calibration and self-regulated learning. *Educational Psychology Review*, *12*(4), 437–475.

Tanner, K. (2012). Promoting student metacognition. *CBE—Life Sciences Education*, *11*(2), 113–120. doi: 10.1187/cbe.12-03-0033

van Loon, M. H., de Bruin, A. B. H., van Gog, T., & van Merrienboer, J. J. G. (2013). Activation of inaccurate prior knowledge affects primary-school students' metacognitive judgments and calibration. *Learning and Instruction*, *24*, 15–25. doi: 10.1016/j.learninstruc.2012.08.005

Winne, P. H., & Hadwin, A. R. (1998). Studying as self-regulated learning. In Hacker, D. J., Dunlosky, J., & Graesser, A. C.(Eds.), *Metacognition in educational theory and practice* (pp. 277–304). New York: Routledge.

Wright, S. L., Jenkins-Guarnieri, M. A., & Murdock, J. L. (2013). Career development among first-year college students; self-efficacy, student persistence, and academic success. *Journal of Career Development*, *40*(4), 292–310. doi: 10.1177/0894845312455509

Ziegler, B., & Montplaisir, L. (2014). Student perceived and determined knowledge of biology concepts in an upper-level biology course. *CBE—Life Sciences Education*, *13*(2), 322–330. doi: 10.1187/cbe.13-09-0175

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory into Practice*, *41*(2), 64–70. doi: 10.1207/s15430421tip4102_2

Zimmerman, B. J., Moylan, A., Hudesman, J., White, N., & Flugman, B. (2011). Enhancing self-reflection and mathematics achievement of at-risk urban technical college students. *Psychological Test and Assessment Modeling*, *53*, 108–127.

Zohar, A., & David, A. B. (2009). Paving a clear path in a thick forest: A conceptual analysis of a metacognitive component. *Metacognition and Learning*, *4*(3), 177–195. doi: 10.1007/s11409-009-9044-6

Zusho, A., Pintrich, P. R., & Coppola, B. (2003). Skill and will: The role of motivation and cognition in the learning of college chemistry. *International Journal of Science Education*, *25*(9), 1081–1094. doi: 10.1080/0950069032000052207