Check for updates

Received: 15 July 2020 | Revised: 17 March 2021 | Accepted: 17 May 2021

DOI: 10.1002/bmb.21547

ARTICLE

Biochemistry and Molecular Biology Education

WILEY

Reevaluate how to evaluate: Systemic assessment biases affect students' confidence in college upper-division biology laboratory courses

Kesten Bozinovic[1,2] | **Zuying Feng**[2] | **Christine M. Stewart**[1,2] | **Darcy C. Engelhart**[2,3] | **Sylvia Gong**[1,2] | **Jeanne P. Vu**[1,2] | **Miguel F. Vasquez**[2,4] | **Goran Bozinovic**[2,3]

[1]Extended Studies, University of California San Diego, La Jolla, California, USA

[2]Boz Life Science Research and Teaching Institute, San Diego, California, USA

[3]Division of Biological Sciences, University of California San Diego, La Jolla, California, USA

[4]Department of Cellular and Developmental Biology, University of California San Diego, La Jolla, California, USA

Correspondence
Goran Bozinovic, Division of Biological Sciences, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093.
Email: gbozinovic@ucsd.edu

Abstract

Grades influence students' confidence and decisions to complete STEM degrees and pursue relevant careers. What affects students' confidence and performance in college upper-division biology laboratory courses and how relevant are evaluation methods to career success? STEM laboratory courses are an excellent model to address these issues because of the hybrid environment, combining traditional lecture course format and the practical application of knowledge. We surveyed 567 students in two upper-division laboratory molecular biology courses at a major research university to compare course-content self-assessment, students' predicted grades, and actual grades received. By analyzing students' confidence and correlating them to grades assigned by the instructor, we identified biases including student and Instructor Assistant (IA) gender, IA experience, and academic quarter. Considering the systemic effect of identified biases, a correlation ($R^2 = 0.37$, $p < 0.01$) between predicted and actual grades, and weak but statistically significant correlation ($R^2 = 0.10$, $p < 0.01$) between students' comprehensive course-content self-assessment and their predicted grade are not surprising. Our analysis suggests that students' quantifiable self-assessment, a relatively simple and data-rich resource, helps identify evaluation bias. If administered periodically throughout the course, these assessments can help mitigate biases, improve student learning, evaluation, and retention in STEM fields.

KEYWORDS

bias, biochemistry, laboratory, recombinant DNA, self-assessment, self-efficacy, STEM, undergraduate

1 | INTRODUCTION

Should science, technology, engineering, and mathematics (STEM) college students be evaluated on their aptitude and potential for successful transitions into life science careers including graduate school, professional programs, and the workforce? Course performances are distilled into a grade point average (GPA) value, a significant filter in graduate and professional school applications, resulting in a competitive and bottom-line model

 © 2021 International Union of Biochemistry and Molecular Biology

within the learning environment. College grades should reflect subject mastery and practical proficiency, which influence students' confidence and perception of their abilities.[1] Grades influence students' decision to complete STEM degrees and pursue relevant careers, so evaluation biases between and within STEM courses should be evaluated.[2] While grades are a formal record of students' competence, students' confidence does not always reflect the grade received.[3] Students receiving lower grades are likely less confident in future courses, perpetuating a cycle of lower grades. Challenging and low average grade courses contribute to a gender gap in STEM because women tend to place a higher value on grades.[4] Consequently, fewer women than men with STEM degrees choose STEM careers,[5] suggesting that grades influence career choices and STEM workforce demographics.

Since GPA is often correlated to hiring offers, salaries, and career success,[6] grading policies should mitigate biases in college courses where assessment metrics, including exam scores, grades received, and evaluation of instructor's effectiveness, negatively correlate to large class sizes.[7,8] Teaching and evaluation in large enrollment classes amplify implicitly biased grading and leave students unprepared for STEM careers. Many large enrollment courses are divided into discussion sections taught by Instructional Assistants (IAs). Curriculum standardization and grading normalization are difficult to maintain among sections in general and more difficult considering variation in IA leniency.[9] Instructors often adjust grading to maintain normally distributed grades; while this approach may mitigate the instructor's bias, it is based on peer performance. A student among better-performing peers will receive a lower grade than a student among under-performing peers.[1] Moreover, evaluation methods in large-enrollment courses are often based on memorization[10] and provide a cross-section of students' performance, often irrelevant to career success.[11–13] Students evaluated on a standardized grading system receive grades as external motivators for their performance. When coupled with memory-based evaluations, this rewards system detracts from students' intrinsic motivation and often negates the very creativity and problem-solving that scientific endeavor is founded upon.[13,14] STEM undergraduates are unprepared for the workplace because they lack comprehensive practical and soft skill sets.[15–17] Life-science undergraduate laboratory courses should rely on evaluation models similar to graduate programs, promoting continuous feedback and opportunity for trial and error to better prepare students for the transition to science professions. Evaluations should quantify students' mastery of concepts, a factor that influences students' perception of their abilities. Academic self-efficacy is a moderate predictor of student course success,[18] so by considering students' confidence of concept mastery and their performance assessment, instructors can promote an intrinsic motivation to learn.

The limiting factor in large enrollment courses is the structure; high student-to-instructor ratios make individualized instruction and evaluation difficult. The standard STEM laboratory course is a hybrid environment, combining the disadvantages of a large undergraduate course in lecture settings, and the advantages of a personalized, smaller student-to-instructor ratio within lab sections. Lab sections provide an engaging experience where students work in small groups, applying concepts with feedback from the instructor and IAs. The lecture/lab contrast of a traditional college laboratory course is an excellent model to identify systemic biases affecting both grades and students' confidence.

We administered end-of-course assessments to quantify students' confidence of course content and overall performance, then compared each to the final grade assigned by the instructor. The assessments included content-specific theoretical, practical, communication, personal, and "big-picture" sections, and a student's prediction of their final grade. This format is more reflective of course effectiveness than generic end-of-course science department issued surveys.[19] We hypothesized that students' performance and confidence in upper-division molecular biology laboratory courses positively correlate but are affected by systemic biases. We propose a relatively simple, course-specific, data-rich model for student evaluation that could help identify and mitigate systemic evaluation biases, improve student engagement, and promote transitional learning focused on critical and practical skill development.

2 | METHODS

2.1 | Design

Students' self-assessment data was collected ($n = 567$) at the end of two upper-division biology laboratory courses at the University of California San Diego (UCSD) over eight academic quarters from 2016 to 2018. Four hundred and fifty-two (452) students (55% females and 45% males) were from "Biochemical Techniques" (BC), which emphasizes translational biology: protein purification, expression, identification and quantification including size exclusion and affinity column chromatography, western blot/ELISA, fluorescent protein expression vectors, and bioinformatics. One hundred and fifteen (115) students (57% females and 43% males) were from "Recombinant DNA Techniques" (RD), which emphasizes transcriptional biology and gene expression,

TABLE 1 Survey response breakdown of two upper-division molecular biology laboratory courses by relevant academic calendar of student participants per gender

Course	Year	Quarter	Female students	Male students	Total
Biochemical Techniques (BC)	2016	Winter	72	44	116
		Spring	62	60	122
	2017	Winter	60	52	112
		Summer Session I	18	8	26
		Summer Session II	15	16	31
	2018	Winter	23	22	45
	Total		**250**	**202**	**452**
Recombinant DNA Techniques (RD)	2016	Summer Session II	25	13	38
	2017	Fall	41	36	77
	Total		**66**	**49**	**115**
Overall Total			**316**	**251**	**567**

including nucleic acid quantification, gel electrophoresis, mutagenesis and cloning, PCR/qRT-PCR/RNAi, and bioinformatics (Table 1). Both courses have the same prerequisite: "The Cell," an introduction to cellular structure and function.

An academic year spans a 10-week fall, winter, and spring quarter. Students can take either course during the accelerated five-week summer session; the course content and laboratory activities during the 10-week and accelerated five-week module are the same. Both courses are taught throughout the year. All students assessed in this study were taught by the same instructor with over 8 years of experience teaching each course. The courses' structure includes three 50-min lectures and two 4-h laboratory sessions per week; lab sections were taught by an MS or PhD student Instructional Assistant (IA) with varying experience. The IA selection criteria include full-time graduate standing and >3.0 GPA in upper-division and graduate coursework. Preference was given to students who completed the course with an A. New IAs completed a graduate "Introduction to College Biology Education" apprenticeship course.

Students were assessed throughout BC and RD with three in-class exams, two comprehensive lab reports formatted per peer-review manuscript standards, four quizzes administered by IAs, four lab notebook checks, and extra credit activities. Exam questions were designed by the instructor and graded by IAs using an instructor-provided answer key. Extra credit (approximately 2% of the final grade) included in-class experimental design activities and a cover letter and a one-page resume highlighting students' course relevant practical skills (>95% of the students completed the extra credit requirements). Prior to the surveys, students received feedback on all assignments except the second lab report and final exam (approximately 30% of their final grade). Surveys were used to quantify students' confidence in course theoretical concepts and technical proficiency on a 0 to 5 scale, and students were asked to predict their final grade at the end of the survey (Figures S1 and S2). Upon the completion of the course, students received a single final grade, which reflects the overall performance in both lecture and lab portions of the course.

2.2 | Survey content and administration

Surveys focused on attitude and self-perceived content mastery consisting of 70 questions for RD and 62 questions for BC were administered to students during the final exam with no time constraint. Students were informed that the survey data would not affect their final grade and that their participation in the survey study was anonymous and voluntary. Most questions consisted of topics, skills, and statements reflecting course learning objectives within five categories: theoretical, practical, science communication, personal, and "big picture" (Tables 2 and 3). Students were instructed to rate their proficiency on a Likert scale 0–5, 0 indicating no knowledge or competency ("no clue"), and 5 indicating mastery ("expert"). The last survey questions read: "Based on your answers to the questions above, what grade would you assign yourself in the course?" For this question, students assigned themselves a letter grade within the A to C− range.

2.3 | Data management

Students' names were replaced by random number identifiers, and gender and academic status were kept as

TABLE 2 Self-assessment survey breakdown including the number of questions and a sample statement per category for recombinant DNA techniques (RD) upper-division molecular biology laboratory courses

Category	# of Qs	Sample statement
Theoretical: Transcription biology	23	Tissue-specific gene expression
Practical skills: Experimental design & techniques	20	Biological versus Technical Replicates
Science communication	10	Understand and communicate the limits of your experimental design
Personal	5	Lab seeking opportunities—be specific: know their science and explain why you benefit/why the lab benefits
The big picture: Long term	12	Decision-making: Where does the DATA come from and HOW is the data analyzed?

TABLE 3 Self-assessment survey breakdown including the number of questions and a sample statement per category for biochemical techniques (BC) upper-division molecular biology laboratory courses

Category	# of Qs	Sample statement
Theoretical: Protein biology	18	Cori cycle and substrate affinity
Practical skills: Experimental design & techniques	19	Protein quantification; loading control/ housekeeping genes
Science communication	10	Understand and communicate the limits of your experimental design
Personal	5	Lab seeking opportunities—be specific: know their science and explain why you benefit/why the lab benefits
The Big Picture: Long term	10	Honor & pay attention to individual biological variation

variables for the analysis. The course instructor had no access to survey data until it was entered by the project student assistant 2 weeks after the final grades were submitted to university records.

Three data categories were collected: survey data, categorical data, and IA information. Student information accessed from the university's records and course grade sheets included the final course grade, gender, class level (freshman, sophomore, etc.), quarter, and year of the lab course. IA information (gender and overall/course-specific teaching experience) was used to identify biases. IAs were ranked as "inexperienced" (0) if they had not taught either course, "experienced" (1) if they taught an alternative course once, and "very experienced" (2) if they taught the same course once (Table 4).

Students' confidence scores were averaged and normalized to standard GPA values (ex. $A = 4.0$, $A- = 3.7$, $B+ = 3.3$, etc.). Confidence scores 4.0–5.0 were considered as A's, 3.0–3.9 as B's, 2.0–2.9 as C's, 1.0–1.9 as D's, and 0–1.0 as F's. Plus (+) and minus (−) marks were considered within the top and bottom 20% of each numeric range. Per the syllabi, scores between 1.9 and 1.0 were not assigned plus or minus marks. Normalized mean confidence scores and student-predicted grades were compared to the instructor's end-of-course assigned grade.

To evaluate students' prediction accuracy, the difference between students' responses to their assigned grades, "confidence scores" were calculated by subtracting survey scores (survey confidence) and predicted grade (prediction confidence) from assigned grades. Positive values indicate overconfidence and negative values indicate under-confidence. Surveys of students who received a failing grade (2%) and incomplete surveys (27 total; 5% in BC and 1% RD) were excluded from the analysis.

2.4 | Statistical analysis

We grouped the data for relevant pairwise comparisons to run one-way ANOVA and post-hoc Bonferroni multiple comparison of means to evaluate comparisons between survey scores and predicted and assigned grades and the following independent variables: course, student gender, IA gender, IA experience, quarter and year, and survey sections. Linear regression least squares fit analysis was used to correlate survey scores and predicted grades with assigned grades, and predicted grades to survey scores. Multiple Correspondence Analysis (MCA) was performed to assess the interactions between the independent variables influencing student confidence in SAS JMP 14.3.0 Statistical Software. Two-way hierarchical clustering Fast Ward minimum variance analysis of survey data for each question was performed per course in SAS JMP Genomic 9.1. All statistical analyses, excluding the two-way hierarchical clustering and MCA, were

TABLE 4 Survey response breakdown of two upper-division molecular biology laboratory courses by relevant academic calendar of instructional assistants per gender and experience. IA experience is represented as "0" for IAs who taught the class once, "1" for IAs who taught a different lab course previously, and "2" for IAs who taught the same class twice

			IA gender & experience						Total IAs
Course	**Year**	**Quarter**	**F0**	**M0**	**F1**	**M1**	**F2**	**M2**	**per quarter**
Biochemical Techniques (BC)	2016	Winter	2	4					**6**
		Spring	2				2	2	**6**
	2017	Winter	5	1					**6**
		Summer Session I		2					**2**
		Summer Session II	1					1	**2**
	2018	Winter	1	1		1			**3**
	Total		**11**	**8**		**1**	**2**	**3**	
Recombinant DNA Techniques (RD)	2016	Summer Session II		2					**2**
	2017	Fall	2	1		1			**4**
	Total		**2**	**3**		**1**			
Overall Total			**13**	**11**		**2**	**2**	**3**	

performed using GraphPad Prism 8.4.2, and a nominal p-value cut-off of $p < .05$ was applied.

The study has been reviewed by an IRB and certified as exempt from IRB review under 45 CFR 46.101(b) category 4. The author declares no conflict of interest. There is no specific funding source associated with this study.

3 | RESULTS

3.1 | Survey responses overview

Students and questions are grouped by score range (0–5) for each course by two-way hierarchical clustering (Figure 1). Four clusters in RD courses represent student confidence groups (Figure 1(a)): cluster A is "most confident" (most questions per student were scored between 4–5), cluster B is "moderate confidence" (about half the scores are in the 4–5 range), cluster C is "low confidence" (2–3 score range), and cluster D is "least confident" (0–1 score range). Three confidence clusters within BC courses have a similar distribution (Figure 1(b)): cluster A is "most confident," cluster B is "moderate confidence," and cluster C "least confident."

3.2 | Relationship between survey scores, predicted grades, and assigned grades

Survey and grade prediction data for both courses is normally distributed (D'Augustino-Person Test, RD: $p = 0.993$; $p = 0.311$, respectively; BC: $p = 0.231$; $p = 0.253$, respectively). Predicted grades correlate stronger to assigned grades (Figure 2; $R^2 = 0.37$, $p < 0.01$) than to survey scores ($R^2 = 0.30$, $p < 0.01$); survey scores correlate with assigned grades, but only 10% of the variation is explained by this interaction ($R^2 = 0.1$, $p < 0.01$).

Pairwise comparisons revealed survey scores ($\bar{x}_s = 3.45$) are overall higher than predicted grades ($\bar{x}_p = 3.25$; $p < 0.01$), and both are higher than assigned grades ($\bar{x}_a = 3.16$, $p < 0.01$). RD survey scores ($\bar{x}_{Rs} = 3.46$) and predicted grades ($\bar{x}_{Rp} = 3.26$; $p < 0.01$) are higher than assigned grades ($\bar{x}_{Ra} = 3.00$; $p < 0.01$), while BC survey scores ($\bar{x}_{Bs} = 3.45$) are higher than predicted grades ($\bar{x}_{Bp} = 3.24$; $p < 0.01$) and assigned grades ($\bar{x}_{Ba} = 3.21$; $p < 0.01$). BC students earn higher grades ($\bar{x}_{Ba} = 3.21$) than RD students ($\bar{x}_{Ra} = 3.00$, $p < 0.01$).

Female students are less confident in predicting their grade ($\bar{x}_{fp-fa} = 0.03$), but receive higher grades ($\bar{x}_{fa} = 3.23$) compared to males ($\bar{x}_{mp-ma} = 0.17$, $\bar{x}_{ma} = 3.08$, respectively; $p = 0.043$). Female survey scores ($\bar{x}_{fs} = 3.46$) are higher than their predicted grades ($\bar{x}_{fp} = 3.25$; $p < 0.01$) and assigned grades ($\bar{x}_{fa} = 3.23$; $p < 0.01$). Likewise, male survey scores ($\bar{x}_{ms} = 3.44$) are higher than their predicted grades ($\bar{x}_{mp} = 3.24$; $p < 0.01$) and assigned grades ($\bar{x}_{ma} = 3.08$; $p < 0.01$), while predicted grades are higher than their actual grades ($p = 0.03$).

Males in BC ($\bar{x}_{Bma} = 3.15$) receive higher grades than males in RD (Figure 3(a); $\bar{x}_{Rma} = 2.80$, $p < 0.01$). RD female students have higher survey scores ($\bar{x}_{Rfs} = 3.50$) than assigned grades ($\bar{x}_{Rfa} = 3.10$; $p < 0.01$), while RD male students' survey scores ($\bar{x}_{Rms} = 3.42$) and predicted grades ($\bar{x}_{Rmp} = 3.24$; $p < 0.01$) are higher than assigned

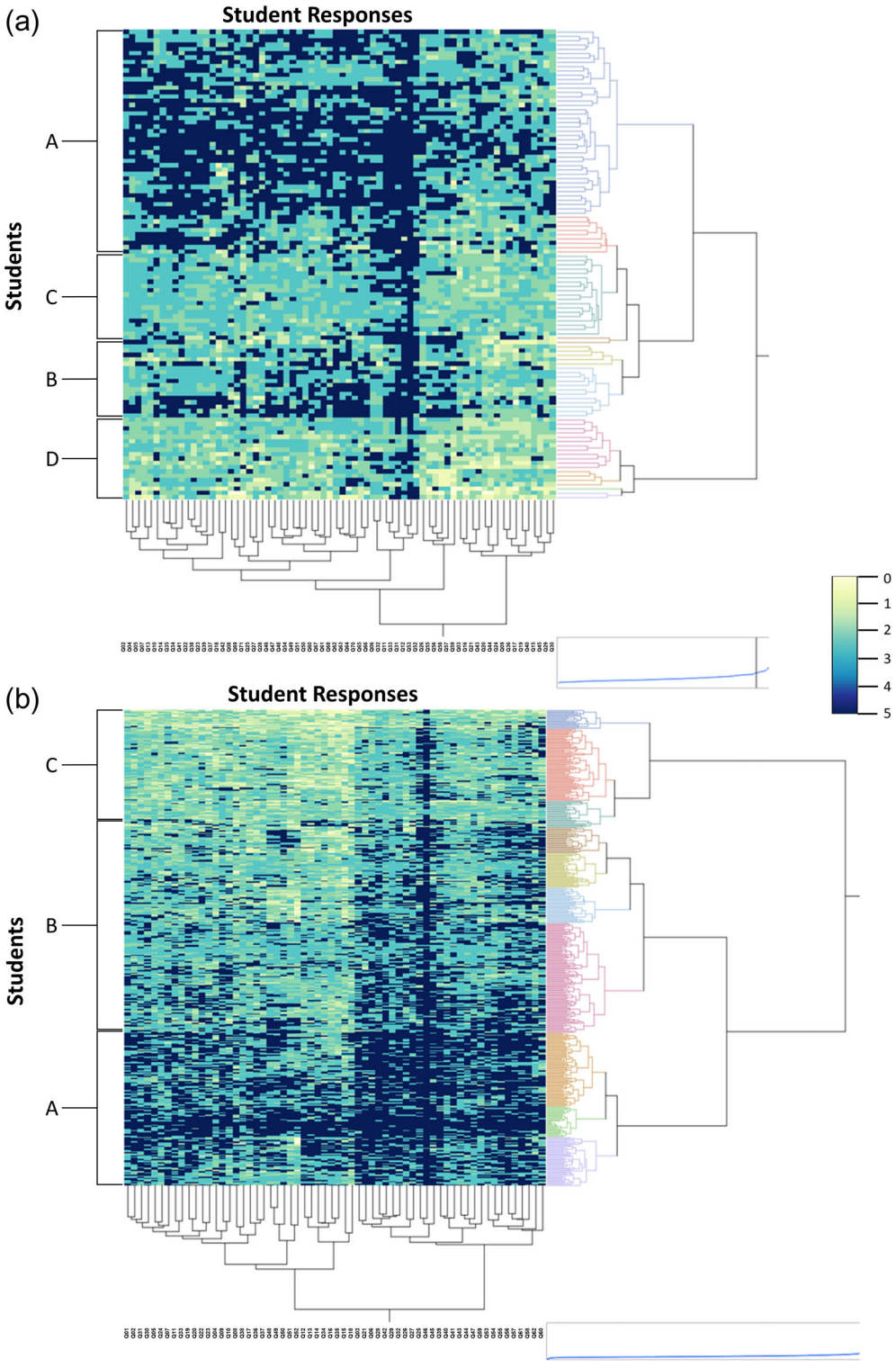

FIGURE 1 Student responses to survey questions in two upper-division molecular biology laboratory courses. *2-way* hierarchical clustering of individual student responses to survey questions by course. Data are grouped by students and questions, quantified as student responses on a 0–5 scale. Light yellow indicates low scores, and dark blue indicates high scores. (a) Heatmap of student responses from recombinant DNA techniques (RD) course ($n = 115$). (b) Heatmap of student responses from biochemical techniques (BC) course ($n = 452$)

grades ($\bar{x}_{Rma} = 2.80$; $p < 0.01$). BC females' and males' survey scores ($\bar{x}_{Bfs} = 3.44$; $\bar{x}_{Bms} = 3.45$, respectively) are higher than both predicted grades ($\bar{x}_{Bfp} = 3.24$; $\bar{x}_{Bmp} = 3.24$, respectively; $p < 0.01$) and assigned grades ($\bar{x}_{Bfa} = 3.26$, $p = 0.03$; $\bar{x}_{Bma} = 3.15$; $p < 0.01$, respectively).

Prediction confidence scores increase as assigned grades decrease, except for students who earned Cs (Figure 3(b), Table S1): "A students" have lower prediction confidence scores ($\bar{x}_{Ap-Aa} = -0.29$) than students who earn B+ and below grades ($p < 0.01$). "A− students" are less confident in predicting grades ($\bar{x}_{A^-p-A^-a} = -0.13$) than "B and below students" ($p = 0.014$). "B+ students" ($\bar{x}_{B^+p-B^+a} = 0.05$) are less confident than "B− and below students" ($p = 0.01$), while "B and B− students"

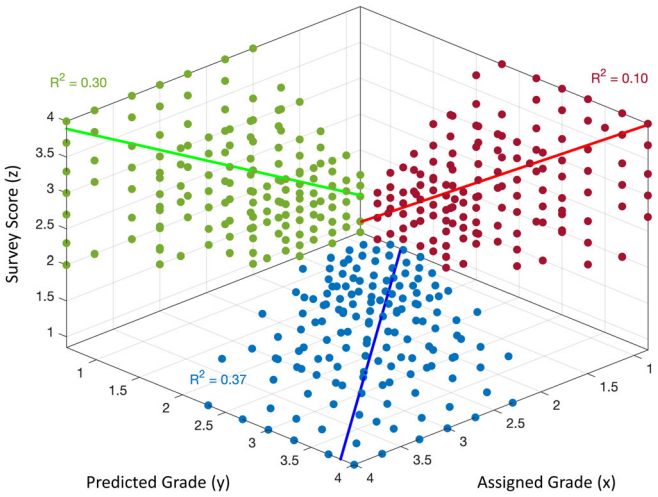

FIGURE 2 Predicted grades have a stronger positive correlation to assigned grades than students' survey scores. Scatterplot of "predicted grade," "survey score" and "assigned grade." Correlation of predicted grade and assigned grade (blue; $R^2 = 0.37$) and survey scores to assigned grade (red; $R^2 = 0.10$). Correlation of predicted grade to survey scores (green line; $R^2 = 0.30$; linear regression least squares fit line analysis, $p < 0.01$)

($\bar{x}_{Bp-Ba} = 0.12$, $\bar{x}_{B^-p-B^-a} = 0.28$) are less confident than "C+ students" ($\bar{x}_{C^+p-C^+a} = 0.56$; $p < 0.01$).

The grade prediction confidence pattern is the same as the confidence scores pattern: all survey confidence scores increase as grades decrease, except for students who earn Cs (Figure 3(c), Table S2). "A and A− students" are the only group that underpredicts their survey scores and predicted grades.

3.3 | IA effect on Students' confidence

IA gender has a greater positive effect on male students' overconfidence than on female students (Figure 4(a)). Male students with male IAs significantly overpredict their grades ($\bar{x}_{mmp-mma} = 0.24$) compared to female students with male IAs ($\bar{x}_{fmp-fma} = 0.04$; $p = 0.02$) and female students with female IAs ($\bar{x}_{ffp-ffa} = 0.007$, $p < 0.01$). Students with more experienced IAs are significantly underconfident ($\bar{x}_{e2p-e2a} = -0.15$) and underpredict their grade compared to students with inexperienced IAs teaching the course for the first time (Figure 4(b); $\bar{x}_{e0p-e0a} = 0.10$; $p < 0.01$). Students with slightly more experienced IA's who taught a different lab course are overconfident in their grade prediction ($\bar{x}_{e1p-e1a} = 0.16$), compared to students with experienced IAs ($p < 0.01$). Only students with IAs teaching the course for a second time underpredict their grades.

3.4 | Effect of academic quarter on student confidence

Students in each quarter generally overpredict their survey scores (Figure 5(a)). Students were most overconfident during the Fall quarter and received the lowest grades ($\bar{x}_{Fa} = 2.86$) compared to aggregate Winter ($\bar{x}_{WIa} = 3.24$; $p < 0.01$), Spring ($\bar{x}_{SPa} = 3.20$; $p < 0.01$), and Summer (Figure 5(b); $\bar{x}_{SSa} = 3.14$, $p < 0.05$) quarters. Fall 2017 grade prediction confidence ($\bar{x}_{F7p-F7a} = 0.33$) is significantly higher than the Winter 2017 ($\bar{x}_{W7p-W7a} = 0.08$; $p = 0.03$) and Spring 2016 (Figure 5(a); $\bar{x}_{S6p-S6a} = -0.11$, $p < 0.01$) quarters. Only students in Spring 2016 underpredict their grades ($\bar{x}_{S6p-S6a} = -0.11$).

3.5 | Survey section analysis

Students are more confident in their communication and big picture related skills than in the other sections in the survey (Tables S3 and S4). RD "Communication" scores ($\bar{x}_{RCO} = 3.31$) are significantly higher than "Theoretical" ($\bar{x}_{RTH} = 3.09$; $p < 0.05$), and "Personal" ($\bar{x}_{RPE} = 3.02$; $p < 0.01$) scores. The average survey scores for the "Big Picture" section ($\bar{x}_{RBI} = 3.27$) are higher than the "Personal" section ($\bar{x}_{RPE} = 3.02$, $p < 0.01$). Likewise, BC "Communication" scores ($\bar{x}_{BCO} = 3.32$) are significantly higher than the "Practical" ($\bar{x}_{BPR} = 3.12$; $p < 0.01$), "Theoretical" ($\bar{x}_{BTH} = 3.01$; $p < 0.01$), and "Personal" ($\bar{x}_{BPE} = 2.92$; $p < 0.01$) sections. The "Big Picture" survey scores ($\bar{x}_{BBI} = 3.35$) are higher than both the "Practical" ($p < 0.01$) and "Theoretical" ($p < .01$) sections, while "Practical" scores are higher than the "Personal" ($p < 0.01$) and "Theoretical" scores ($p < 0.01$). The highest scored sections are not predictors of higher grades received in the course.

3.6 | Effect of survey questions on students confidence

RD students are most confident (score = 5) in "Do Not take credit for someone else's work—Do not plagiarize!" (94%), "gel electrophoresis" (75%), "Good Laboratory Practices (GLP)—how to keep your lab notebook and lab safety" (70%), "DNA, RNA extraction, purification" (58%), and "Citations: reference sources EACH time you mention something that is not a common knowledge" (58%) statements. RD students are least confident in "Grad school/Med school applications," "genomic and cDNA libraries," and "expression vectors." When analyzed by gender, similar low and high confidence patterns were observed for male students and high

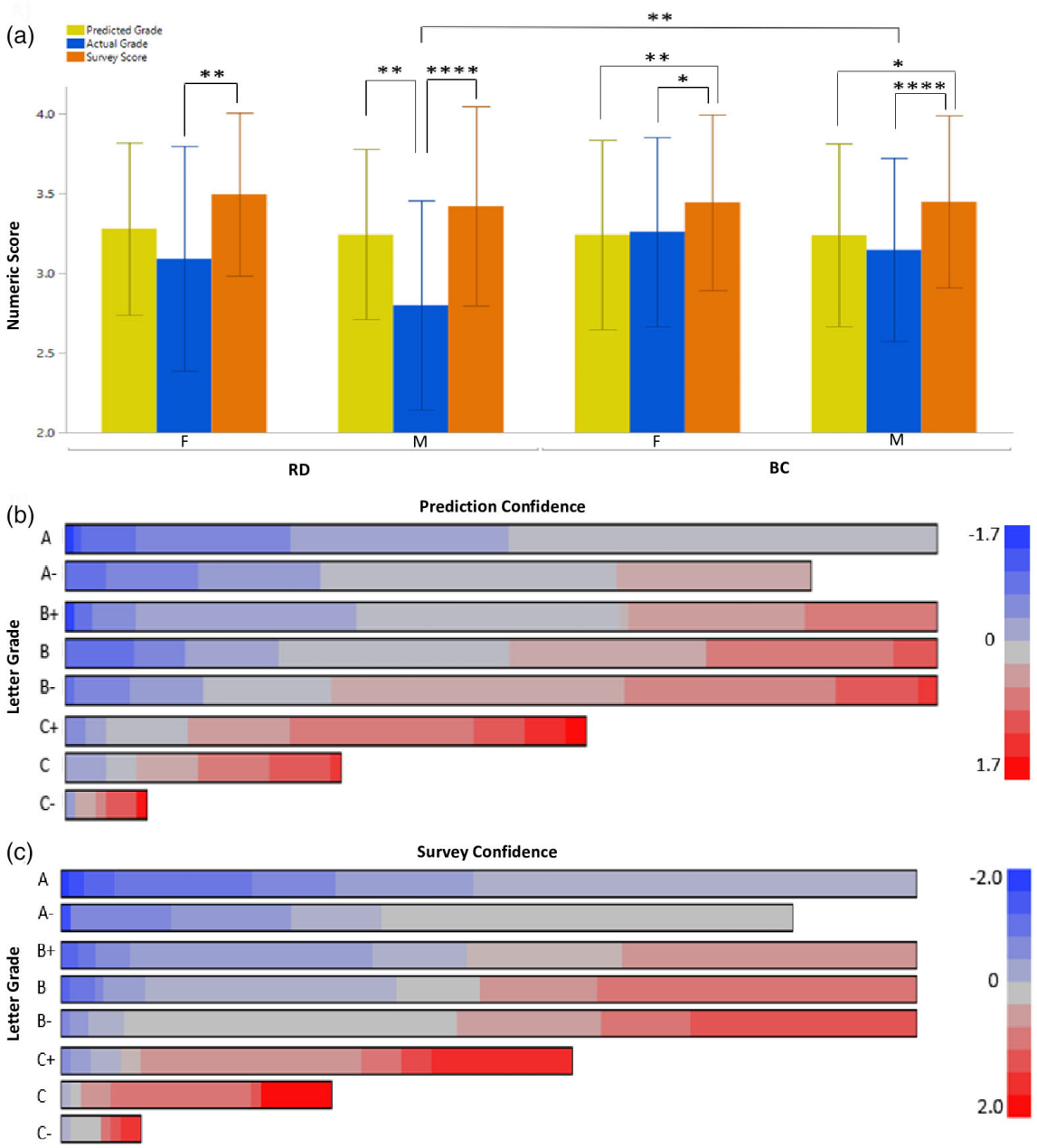

FIGURE 3 Relationship between survey scores, predicted grade, and assigned grade by course and gender; students who earn high grades (A and A−) tend to underpredict, while students who earn lower grades (B+ and lower) tend to overpredict their actual grade (ANOVA Bonferroni correction, $p < 0.05$). Statistical significance is denoted as * for $p \leq 0.05$, ** for $p \leq 0.01$, *** for $p \leq 0.001$, and **** for $p \leq 0.0001$. (a) Students' mean survey scores (orange bars) are significantly higher than their assigned grades (blue) for female ($n = 316$) and male ($n = 251$) students in both recombinant DNA techniques (RD, $n = 115$) and biochemical techniques (BC, $n = 452$) laboratory courses; female ($n = 250$) and male ($n = 202$) students in BC course have significantly higher mean survey scores (orange) compared to predicted grades (yellow). Only males in RD ($n = 49$) have significantly high predicted grades (yellow) relative to actual grades (blue). BC males ($n = 202$) earned higher actual grades than RD males ($n = 49$) (ANOVA Bonferroni correction, $p < 0.05$). (b) Cell plot of student "prediction confidence" scores based on the grade prediction at the end of the administered survey. (c) Cell plot of student "survey confidence" mean scores derived from 70 survey questions (RD) and 62 survey questions (BC) per student. Confidence scores (b–c) are calculated as the difference of survey score means (red) and predicted grades (blue) to assigned grades. Gray (0) indicates accurate prediction, dark blue (−0.2) indicates underconfidence, and dark red (2.0) indicates overconfidence

confidence pattern for female students. Female RD students are distinctively least confident about experimental design and technical terminology survey questions: "microarrays and RNASeq," "Sample size and replication," "Biological versus Technical Replicates," and "Standard Curve."

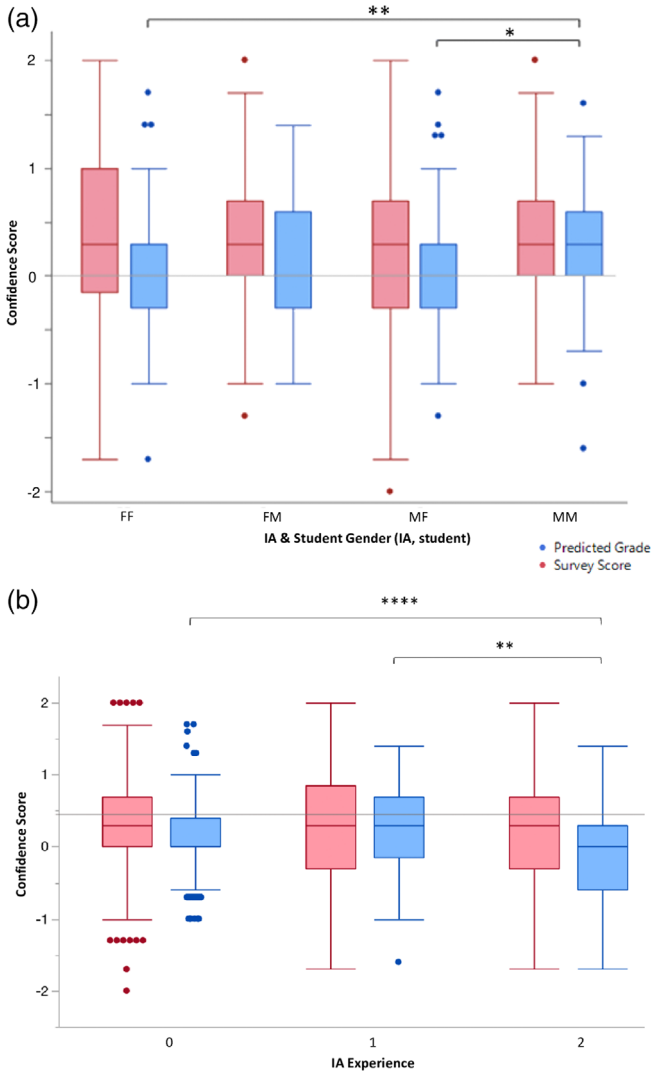

FIGURE 4 Analysis of IA effect on students' confidence: Male students with a male IA are more likely to overestimate their grade compared to female students; students with more experienced IAs underpredict their assigned grade (ANOVA Bonferroni correction, $p < 0.05$). Statistical significance is denoted as * for $p \leq 0.05$, ** for $p \leq 0.01$, *** for $p \leq 0.001$, and **** for $p \leq 0.0001$. Confidence scores are calculated as the difference of survey score means (red) and predicted grades (blue) to assigned grades. Positive values on *Y*-axis represent students' overconfidence, negative values represent underconfidence. (a) Analysis of IA and student gender: Female IA and female student (FF, $n = 153$), female IA and male student (FM, $n = 150$), male IA and female student (MF, $n = 163$), and male IA and male student (MM, $n = 101$). MM students' prediction confidence scores (blue bar) are higher than MF and FF students. (b) IA experience represented as three categories based on teaching experience; "0" represents students' data from an IA's section who taught the class once ($n = 439$), "2" is students' data of sections with IAs who taught the same class twice ($n = 95$), and "1" is students' data from sections with IAs who taught a different lab course previously ($n = 33$). Student data from "0" and "1" IAs have lower predicted confidence scores (blue) relative to "2" IAs

Similarly, female and male BC students are most confident in "do not to plagiarize" (91%), "GLP" (61%), "Healthy Skepticism" (57%), "Negative Data = valuable data" (57%), and "cite properly" (56%) statements. Notably, males are more confident than females about "Column Chromatography: size exclusion, affinity" technical skills (57%), and "Total Protein versus Active Protein" protein quantification concept (53%). Both female and male students are least confident about technical terminology "genomic and cDNA libraries," "bacteria, Restriction Enzymes, methylation," "expression vectors" and the medical/graduate school application process.

3.7 | Overview of the influence of variables on student confidence

Multiple correspondence analysis shows the distribution and clustering of variables mostly defined by grades (Figure 6). Relative to Dimension 1: C+, B+, and A fill the left half of the MCA plot; C and B are in the central region; C−, B−, and A− take the rightmost section. Relative to Dimension 2: C+, C, and C− mark the top section and cluster with high confidence levels; B+, B, and B− occupy the middle of the plot and cluster with median confidence levels; A and A− hold the bottom half of the plot and cluster with low confidence levels. The other variables (course, quarter, gender, IA gender, and IA experience) are closer to the center and have relatively lesser effects on the distribution of confidence levels.

4 | DISCUSSION

Grades impact students' confidence, motivation to learn, decisions to complete STEM degrees, and pursuit of relevant careers.[1,20] It is critical to investigate the effectiveness of student evaluation and identify biases affecting students' grades and confidence. Large enrollment upper-division biology laboratory courses are an excellent model to identify and analyze systemic biases because such courses combine a large, non-personalized lecture structure with the smaller lab setting promoting hands-on activities, practical concept application, and collaboration supported by instructor and IAs. We chose two upper-division college biology laboratory courses with lectures accommodating 48 and 144 students, respectively, while lab sections had an enrollment of 24 students for each course. The courses' structure is favorable to identify biases for several reasons; the student sample size has reasonable statistical power and is sufficient to anticipate normal distribution; female to male distribution is not skewed, while the homogeneous academic status and experience minimize

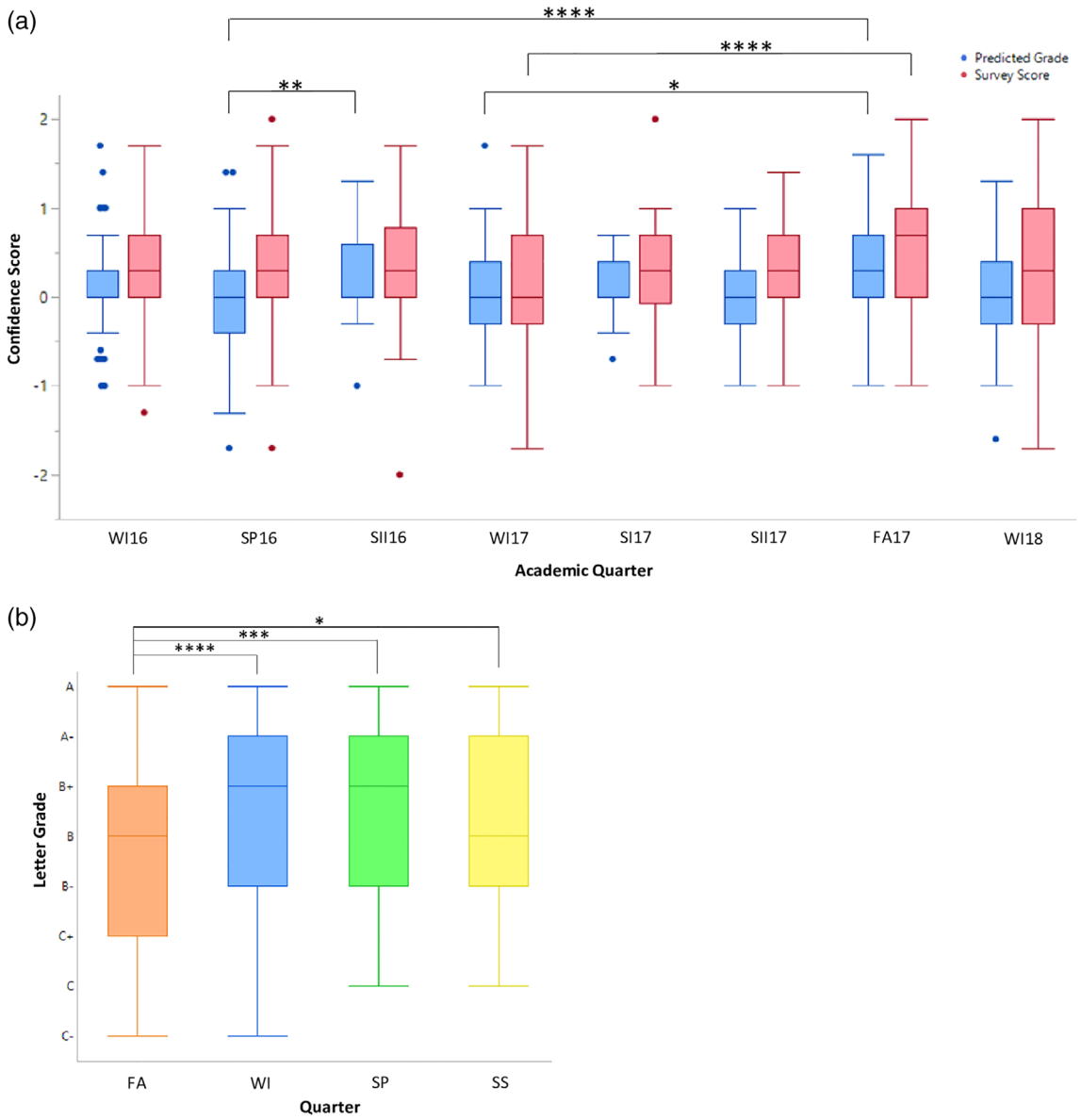

FIGURE 5 Analysis of how academic quarter and year affects students' confidence and assigned grade: Students in fall 2017 most overpredict their assigned grades while earning the lowest grades (ANOVA Bonferroni correction, $p < 0.05$). Statistical significance is denoted as * for $p \leq 0.05$, ** for $p \leq 0.01$, *** for $p \leq 0.001$, and **** for $p \leq 0.0001$. (a) Confidence scores are calculated as the difference of survey score means (red) and predicted grades (blue) to assigned grades. Positive values on *Y*-axis represent students' overconfidence, negative values represent underconfidence. Quarter types assessed include fall (FA), winter (WI), spring (SP), summer session I (SI), and summer session II (SII). Academic quarters, including the year, are represented chronologically along the X-axis: WI16 ($n = 116$), SP16 ($n = 122$), SII16 ($n = 38$), WI17 ($n = 112$), SI17 ($n = 26$), SII17 ($n = 31$), FA17 ($n = 77$), and WI18 ($n = 45$). FA17 ($n = 77$) prediction confidence scores (blue) are significantly higher than WI16 ($n = 116$) scores. FA17 and SII16 ($n = 38$) prediction confidence scores (blue) are also higher than SP16 scores. FA17 survey confidence scores (red) are higher than WI17 scores. (b) Analysis of students' performance (actual grade) by aggregate fall (FA, $n = 77$), winter (WI, $n = 273$), spring (SP, $n = 122$), and summer session (SS, $n = 95$) quarters. FA students earn lower grades than WI, SP, and SS students

confounding factors. IAs can support individual students or smaller groups in real-time; students are evaluated more frequently on professionally-relevant benchmarks based on critical thinking and problem solving including documenting, executing, and troubleshooting experiments, reading and understanding peer-reviewed literature, and writing journal-formatted science reports. Since theoretical lecture concepts are reinforced practically in the lab, we were able to devise a comprehensive content-specific self-assessment.

We correlated students' confidence and assigned grades to identify biases. The two courses were taught by

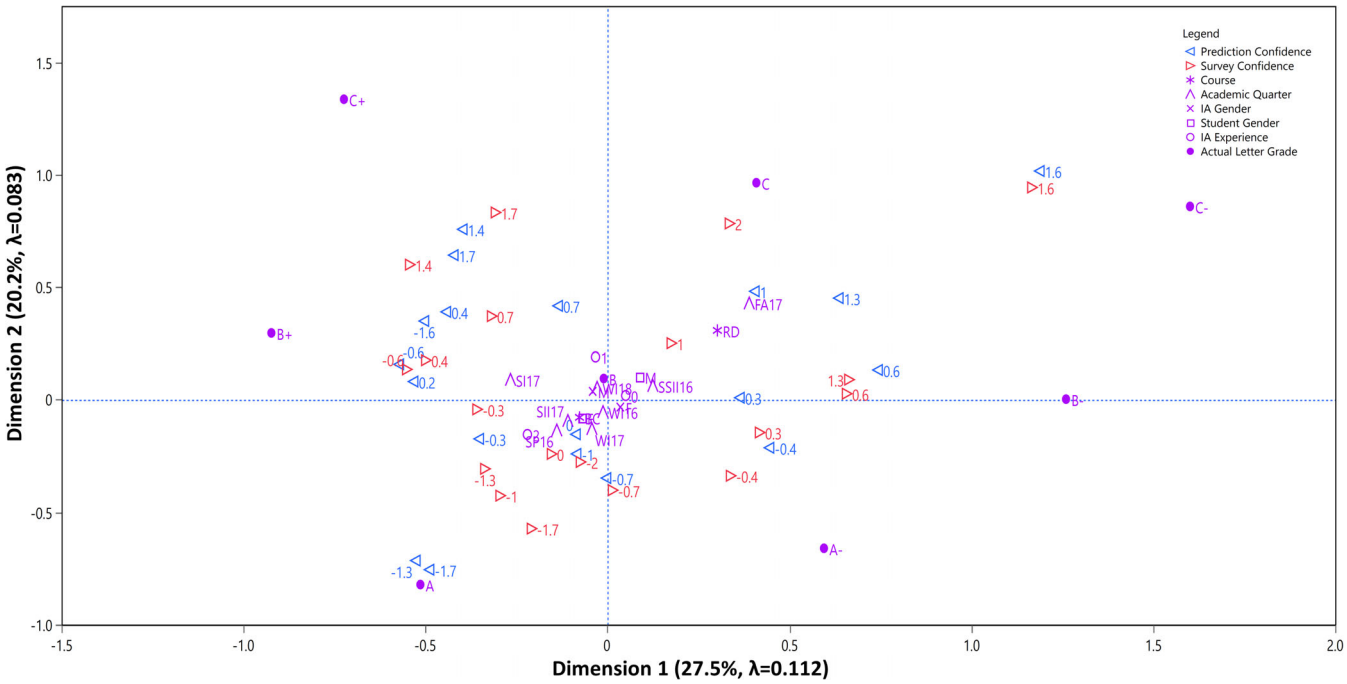

FIGURE 6 Multiple correspondence analysis of variables influencing prediction confidence (blue triangle) and survey confidence (red triangle). Course (star), student gender (box), IA gender (cross), IA experience (circle), academic quarter (caret), and actual letter grade (dot) were considered in the analysis. The distance between two variables corresponds to the strength of the correlation between them: The closer they are, the more closely correlated, and the more likely they are associated with each other. The distance between a variable and the center of the plot corresponds to the strength of that variable's effect on the dataset: The farther a variable is from the center, the more distinct it is from the other variables in the same category, and the more it accounts for the variations in the dataset

the same experienced instructor (>8 years); most of the students were seniors (97% RD and 88% BC; the rest were juniors), so the effects of academic status and students completing both courses with the same instructor were not considered. Students' course evaluation was not directly affected by experimental outcomes to promote trial and error and to minimize students' negative perceptions of their performances and abilities.[21] Students were instructed to predict their course grade based on both their self-assessment and current course evaluation scores, which accounted for 70% of their final grade (two homework assignments, four quizzes, midterm exam and lab report; students did not have feedback on their "mini" lab report and final exam).

The survey control questions were "Good Laboratory Practices" (GLP) as a positive control and "genomic/cDNA libraries" as a negative control. We expected students to be confident about GLP because they could only enroll in lab courses after passing a safety online course, and IAs monitored students carefully to assure GLP during all experiments; GLP was emphasized at the beginning of each lab by the laboratory manager and/or IAs. Moreover, the UC system began implementing rigorous lab safety policies[22] following the death of a lab assistant in 2008. Students' high confidence (4–5 scores) in their lab safety preparedness ranged from 89% in BC to 92% in RD. We expected students to have low confidence about the genomic/cDNA library concept because it was not reinforced by experiments nor mentioned in the lab manual. There was no further course support of this technical concept beyond being introduced in lecture once. Students' low confidence (3–0 scores) ranged from 48% in RD to 58% in BC. Overall, students' confidence is consistent with the expected trends of the positive and negative control questions.

We hypothesized that students' course content confidence, predicted grades, and assigned grades positively correlate but are affected by systemic biases. While normally distributed survey and grade prediction data for both courses do not validate the objectivity of the instructor's grading, they provide the reference for students' perceptions of assigned grades. The survey scores indicate students' perceived capabilities, the assigned grades reflect the students' performance in the course evaluations, and the predicted grades are students' self-assessment based on both their perceived capabilities and performance in the course evaluations. The relationship among these three measurements explains the relative strength of pairwise correlations: the highest correlation is between the predicted grade and the assigned grade

and the lowest correlation is between the survey score and the assigned grade (Figure 2). The low R^2 values indicate there are other factors involved in their correlations, such as individual variation and biases.

Pairwise comparisons were used to determine the effects of student gender, IA gender and experience, and course year and quarter as potential biases. The lack of significant difference in survey scores and predicted grades between genders is likely due to females underpredicting their grades but performing better than males, who are more confident in their grade prediction. Females' perceived competency has been reported lower than males', despite no significant difference in overall academic performance.[23–25] This may contribute to the gender gap in STEM fields, because females are underconfident relevant to their abilities, which affects students' decisions to pursue STEM, despite outperforming the males in our study.[1,5,20,23–25] IA gender further compounds the effect of students' gender on their confidence; male students taught by male IAs are more confident about their course performance and grade prediction than females taught by female IAs (Figure 4(a)). Reportedly, females' confidence improves significantly when both instructor and IAs are female, but if the instructor is male, female students' conceptual understanding does not improve.[26] Since our survey data was collected from male-instructed courses, it is unsurprising that males with male IAs have higher confidence than females with female IAs. Students significantly underpredict grades if they are taught by a more experienced IA (Figure 4(b)). This is relevant because IA experience may exacerbate students' confidence biases if they are not able to adequately support struggling students in the lab setting.

Fall students are more confident but performed worse than students in other academic quarters (Figure 5). Students may be more relaxed entering an academic year after the summer break and enrolling in multiple challenging courses; subsequent physical and mental fatigue could influence their confidence. Students pursuing life science careers and graduate education often seek volunteering and internship opportunities in campus research labs. We did not account for students working in research labs, which may have influenced the data in either direction; students gain more experience and confidence in technical laboratory courses but can become overcommitted and dedicate less time to coursework. It is plausible that fall quarter students are the most lab work inexperienced and unaffected by fatigue in this regard.

MCA summarizes and shows the correlations between confidence levels and variables, and the relative correlation strength (Figure 6). The closer to the center a variable is, the less effect it has on student confidence. Student gender, IA gender and experience, and academic quarter have statistically significant associations with confidence levels as shown in previous analyses, but their being closer to the center suggests that an individual variable has limited effects on confidence, and the combination of variables results in the large variation of confidence among the students. B, the grade in the middle of the letter grade scale, is at the center of the MCA plot, suggesting it has little associations with confidence levels. Among all variables, grades other than B are the farthest from the center, indicating they have the most significant effects on confidence levels.

Overall, better-performing students (A to B+ range) are less confident and underpredict their grades (Figure 3(b–c)). It is likely that students who earn better grades work harder to compensate for their perceived lack of confidence in course material and success. Conversely, students who receive Cs are more confident than A–B students, but confidence levels between C+, C, and C− did not follow the same pattern of increasing confidence as grades decreased. Possibly, the academic effort is less varied from C+ to C−, so students' confidence is not as significantly impacted by their course performance within that range. Generally, lower-performing students tend to be overconfident because they lack awareness of their abilities.[27] Early administered, confidential, periodic self-assessment surveys could help instructors identify if students' confidence correlates to their course performance and provide different tiers of support. Underperforming and low-confidence students need the most immediate intervention, while underperforming and high-confidence students need early intervention to prevent them from lagging in the course. High-performing and low-confidence students also need attention since they may still be discouraged from pursuing STEM careers if they doubt their abilities. Finally, high-performing and high-confidence students need the least urgent intervention but should be made aware of the challenging transition from the undergraduate academic to a professional environment to prevent complacency and overconfidence.

There is a significant overlap between students' high and low confidence in the two courses, indicating that periodically administered self-assessments could provide data to improve course content and communication in real-time by identifying challenging topics and struggling students. In general, students are more confident in concepts reinforced throughout the quarter or performed in the lab most frequently. Notably, students are not confident about technically challenging molecular biology concepts and experimental design, which were only presented during the lectures, and not sufficiently reinforced during laboratory activities. We did not ask students about their career goals, but since most students are not confident in applying to graduate or medical school, we

suspect this means they may also not be confident to continue into STEM careers. The number of students entering life science graduate programs is not increasing at the same rate as students earning undergraduate life science degrees.[28] Anecdotally, more students surveyed are pursuing jobs in life sciences and other related fields immediately post-graduation rather than applying to graduate and medical school. Interestingly, BC male students were more confident about a protein purification method (column chromatography), and quantification of total versus active protein experimental design. These are "mechanically" inclined methods, suggesting that BC males are more confident in their mechanical abilities. It has been reported that males are significantly more interested and confident in engineering disciplines compared to females.[29,30] Since the molecular biology laboratory methodology is not based on engineering concepts, we expect a minimal effect on students' confidence,

College students' self-efficacy, or perceived competence, affects their academic performance and grades.[3,31] Thus, systemic biases influencing students' confidence and course performance should be routinely identified and mitigated. The current, relatively rigid structure of large-enrollment university courses does not promote preemptive bias mitigation, rather resulting in a grade-based filtering system often irrelevant to students' practical abilities and STEM career potential. Even within the upper-division laboratory class structure where critical thinking, frequent assessments, and personalized instruction are utilized, biases are prevalent. A low-enrollment, personalized, project-based learning environment structured with relevant and comprehensive evaluations rather than a single end-of-course grade would yield a solid metric of students' abilities and career potential.[13,32,33] Content-specific, self-assessment surveys cannot eliminate this oversimplified evaluation problem, but they should help identify struggling students and challenging concepts, blunting bias effects in real-time. We recommend utilizing such surveys periodically to identify and mitigate biases affecting student performance and evaluation. Such a relatively simple and data-rich method could help optimize the course format and learning objectives and promote students' engagement. Our data shows students' confidence is content-specific (Figure 1). Based on survey data distribution, areas of students' high and low confidence can be identified, and instructors can adjust their pace and method of teaching accordingly. Moreover, evaluation biases can be identified by a significant deviation between students' confidence of concepts and assessment results. Confidential surveys coupled with relevant assessments, including practical/methods application, experimental design and troubleshooting, data analysis, manuscript reading comprehension, and written and oral communication can help instructors devise adequately timed, personalized student support.

4.1 | Recommendations/suggestions

- Administer periodic, content-specific, self-assessment surveys and relevant assessments to identify evaluation and confidence biases.
- Compare pre- and post-survey scores to evaluate students' progress and overall course effectiveness.
- Structure the course to utilize a timely hierarchical intervention strategy based on survey data and relevant performance scores offering priority-based tiers of support.
- Perform quality control to assure the effectiveness of the intervention.

4.2 | Study limitations, practical applications, and future direction

Our goal was to quantify students' confidence and ability to predict their grades in large enrollment biology laboratory college courses. By correlating both to assigned grades, we identified performance and evaluation biases (students' gender, IA gender and experience, and academic quarter), which are difficult to control for. The complexity of the data is summarized in Figure 6, revealing clustering patterns of variable interactions. Neither strict points scale grading nor grade normalization ("grading on a curve") mitigates these biases and it is unrealistic to apply a course-specific advanced statistical model in lieu of traditional grading strategies; a degree of subjectivity is unavoidable in heterogeneous learning environments based on a single final score (grade) and eventual GPA metrics.

One of the study limitations is that all data were collected from courses taught by the same male instructor. While this normalizes for instructor experience and grading, we are not able to analyze the effect of instructor gender on student confidence. Moreover, the survey was administered at the end of the course, during the final exam when most students are stressed and overwhelmed so their perception is likely exaggerated. A better approach is a survey administered several days after the end of course. Regardless, this type of questionnaire is especially useful if applied periodically during the course, providing quantitative data of students' intellectual comfort and ability to master concepts. The analysis can help identify students who struggle in the course early, highlight concepts that need reinforcement and discussion, and improve content and students' engagement. The lack of a reliable

reference design (i.e., inherent subjectivity and bias persist regardless of the assessment model) does not protect our experimental design from Type I error (false positive conclusion from rejecting a true null hypothesis) or Type II error (false negative conclusion from accepting a false null hypothesis).

University students with inadequate academic preparation are more likely to drop out of STEM majors.[34] Non-traditional STEM learning environments relying on a project and concept-mastery approach rather than grades are interesting models to study. In this format, self-assessment surveys could quantify students' placement and progress. Without grades, alternative evaluation methods could quantify the effectiveness of an arguably more intuitive, collaborative, student-selected, and project-focused structure.[35] Moreover, self-assessment provides an opportunity for students to consider personal content mastery independent of external assessments,[36,37] perpetuating confidence, which is related to students' motivation to learn.[20] The utility of this model is advisable in a pre-college setting, where students are introduced to concepts ranging from basic laboratory practices to job-seeking soft skills, a key aspect of employability.[38] Moreover, students' decisions about their STEM capabilities are formed throughout their introductory courses; exposing students to fundamental science concepts and techniques in a non-threatening environment before college could help with STEM retention efforts.[39] Frequent assessment of our students' self-perception in traditional academic settings should help educators be more effective in motivating, preparing, and retaining students in STEM careers.

5 | CONCLUSION

Higher science education should equip aspiring scientists to cultivate curiosity, objectivity, and evidence-based truth-seeking relying on statistically disciplined experimentation and solid analytical methods. If educators' goals are to help students explore topics as life-long learners based on their curiosity, creativity, and the challenges they face, the mandate is to implement evaluation methods that promote critical thinking, problem-solving, trial and error, communication, and constant feedback based on mutual respect and long-term goals. Our traditional education system, including large-enrollment undergraduate science departments, often promotes the opposite. College science laboratory courses are an opportunity to improve our students' learning priorities and promote a smoother transition to life-science careers. Our goal was to further the conversation about evaluating students more comprehensively and objectively by identifying systemic biases and propose a simple, affordable, and data-rich method that could improve student engagement and confidence in challenging technical courses. We argue that periodic, comprehensive, content-specific student feedback is a useful tool to improve evaluation methods within an inherently limited and arguably outdated education system.

ACKNOWLEDGMENT

We thank Sophie Liu for processing the data.

AUTHOR CONTRIBUTIONS

All authors performed data analysis and revised the manuscript. Kesten Bozinovic, Zuying Feng, Christine Stewart, and Goran Bozinovic wrote the manuscript; Goran Bozinovic performed research.

ORCID

Goran Bozinovic https://orcid.org/0000-0001-8934-7884

REFERENCES

1. Calsamiglia C, Loviglio A. Grading on a curve: when having good peers is not good. Econ Educ Rev. 2019;73:101916.
2. Stinebrickner R, Stinebrickner TR. A major in science? Initial beliefs and final outcomes for college major and dropout. Rev Econ Stud. 2014;81:426–72.
3. Ferla J, Valcke M, Schuyten G. Judgments of self-perceived academic competence and their differential impact on students' achievement motivation, learning approach, and academic performance. Eur J Psychol Educ. 2010;25:519–36.
4. Rask K, Tiefenthaler J. The role of grade sensitivity in explaining the gender imbalance in undergraduate economics. Econ Educ Rev. 2008;27:676–87.
5. Beede David N, Julian Tiffany A, Langdon David, McKittrick George, Khan Beethika, Doms Mark E. Women in STEM: A Gender Gap to Innovation. SSRN Electronic Journal. http://dx.doi.org/10.2139/ssrn.1964782.
6. Popov SV, Bernhardt D. University competition, grading standards, and grade inflation. Econ Inq. 2013;51:1764–78.
7. Bedard K, Kuhn P. Where class size really matters: class size and student ratings of instructor effectiveness. Econ Educ Rev. 2008;27:253–65.
8. de Paola M, Scoppa V. The effects of class size on the achievement of college students. Manchester Sch. 2011;79:1061–79.
9. Glazer N. Formative plus summative assessment in large undergraduate courses: why both? Int J Teach Learn High Educ. 2014;26:276–86.
10. Rissanen A. Student engagement in large classroom: the effect on grades, attendance and student experiences in an undergraduate biology course. Can J Sci Math Technol Educ. 2018;18:136–53.
11. Bretz RD. College grade point average as a predictor of adult success: a meta-analytic review and some additional evidence. Public Pers Manage. 1989;18:11–22.
12. Cohen PA. College grades and adult achievement: a research synthesis. Res High Educ. 1984;20:281–93.
13. van Dierendonck D, van der Gaast E. Goal orientation, academic competences and early career success. Career Dev Int. 2013;18:694–711.

14. Kohn A. Punished by rewards: the trouble with gold stars, incentive plans, A's, praise, and other bribes. Boston, MA: Houghton Mifflin Co; 1993.
15. Jang H. Identifying 21st century STEM competencies using workplace data. J Sci Educ Technol. 2016;25:284–301.
16. Meier RL, Williams MR, Humphreys MA. Refocusing our efforts: assessing non-technical competency gaps. J Eng Educ. 2000;89:377–85.
17. Tang H-L, Lee S, Koh S. Educational gaps as perceived by IS educators: a survey of knowledge and skill requirements. J Comput Inf Syst. 2001;41:76.
18. Honicke T, Broadbent J. The influence of academic self-efficacy on academic performance: a systematic review. Educ Res Rev. 2016;17:63–84.
19. Combs KL, Gibson SK, Hays JM, Saly J, Wendt JT. Enhancing curriculum and delivery: linking assessment to learning objectives. Assess Eval High Educ. 2008;33:87–102.
20. Fenollar P, Román S, Cuestas PJ. University students' academic performance: an integrative conceptual framework and empirical analysis. Br J Educ Psychol. 2007;77:873–91.
21. Graham MJ, Frederick J, Byars-Winston A, Hunter AB, Handelsman J. Increasing persistence of college students in STEM. Science. 2013;341:1455–6.
22. UC Laboratory Safety Policies. n.d. https://blink.ucsd.edu/safety/research-lab/ucop-labsafety.html. Accessed on 1 July 2020.
23. Ehrlinger J, Dunning How D. Chronic self-views influence (and potentially mislead) estimates of performance. J Pers Soc Psychol. 2003;84:5–17.
24. Leslie Larry L, McClure Gregory T, Oaxaca Ronald L. Women and Minorities in Science and Engineering. J Higher Educ. 1998;69(3):239–76. http://dx.doi.org/10.1080/00221546.1998.11775134.
25. MacPhee D, Farro S, Canetto SS. Academic self-efficacy and performance of underrepresented STEM majors: gender, ethnic, and social class patterns. Anal Soc Issues Public Policy. 2013;13:347–69.
26. Cotner S, Ballen C, Brooks DC, Moore R. Instructor gender and student confidence in the sciences: A need for more role models? J. Coll. Sci. Teach. 2011;40:5.
27. Miller TM, Tyler M, Geraci L. Unskilled but aware: reinterpreting overconfidence in low-performing students. J Exp Psychol Learn Mem Cogn. 2011;37:502–6.
28. National Center for Education Statistics. Master's degrees conferred by postsecondary institutions, by field of study: Selected years, 1970–71 through 2017–18; 2019.
29. Su R, Rounds J. All STEM fields are not created equal: people and things interests explain gender disparities across STEM fields. Front Psychol. 2015;6:189.
30. Woodcock A, Bairaktarova D. Gender-biased self-evaluations of first-year engineering students. J Women Minor Sci Eng. 2015;21:255–69.
31. Zimmerman BJ, Bandura A, Martinez-Pons M. Self-motivation for academic attainment: the role of self-efficacy beliefs and personal goal setting. Am Educ Res J. 1992;29:663–76.
32. Light RJ. Making the most of college: Students speak their minds. Cambridge, MA: Harvard University Press; 2004.
33. Smith Karl A, Sheppard Sheri D, Johnson David W, Johnson Roger T. Pedagogies of Engagement: Classroom-Based Practices. J. Eng. Educ. 2005;94(1):87–101. http://dx.doi.org/10.1002/j.2168-9830.2005.tb00831.x.
34. Sithole Alec, Chiyaka Edward T, McCarthy Peter, Mupinga Davison M, Bucklein Brian K, Kibirige Joachim. Student Attraction, Persistence and Retention in STEM Programs: Successes and Continuing Challenges. High. Educ. Stud. 2017;7(1):46. http://dx.doi.org/10.5539/hes.v7n1p46.
35. R. J. Marzano. Transforming classroom grading. Alexandria, VA: Association for Supervision and Curriculum Development; 2000.
36. Clark Formative I. Assessment: assessment is for self-regulated learning. Educ Psychol Rev. 2012;24:205–49.
37. McMillan JH, Hearn J. Student self-assessment: the key to stronger student motivation and higher achievement. Educ Horizons. 2008;87:40–9.
38. Hogan R, Chamorro-Premuzic T, Kaiser RB. Employability and career success: bridging the gap between theory and reality. Ind Organ Psychol. 2013;6:3–16.
39. Ost B. The role of peers and grades in determining major persistence in the sciences. Econ Educ Rev. 2010;29:923–34.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Bozinovic K, Feng Z, Stewart CM, Engelhart DC, Gong S, Vu JP, et al. Reevaluate how to evaluate: Systemic assessment biases affect students' confidence in college upper-division biology laboratory courses. Biochem Mol Biol Educ. 2021;1–15. https://doi.org/10.1002/bmb.21547