frontiers in Education

ORIGINAL RESEARCH
published: 27 September 2019
doi: 10.3389/feduc.2019.00107

Gender Performance Gaps Across Different Assessment Methods and the Underlying Mechanisms: The Case of Incoming Preparation and Test Anxiety

Shima Salehi[1], Sehoya Cotner[2], Samira M. Azarin[3], Erin E. Carlson[4], Michelle Driessen[4], Vivian E. Ferry[3], William Harcombe[5], Suzanne McGaugh[5], Deena Wassenberg[2], Azariah Yonas[2] and Cissy J. Ballen[6]*

[1] Graduate School of Education, Stanford University, Stanford, CA, United States, [2] Department of Biology Teaching and Learning, University of Minnesota, Minneapolis, MN, United States, [3] Department of Chemical Engineering and Materials Science, University of Minnesota, Minneapolis, MN, United States, [4] Department of Chemistry, University of Minnesota, Minneapolis, MN, United States, [5] Department of Ecology, Evolution and Behavior, University of Minnesota, Minneapolis, MN, United States, [6] Department of Biological Sciences, Auburn University, Auburn, AL, United States

OPEN ACCESS

Edited by:
Subramaniam Ramanathan, Nanyang Technological University, Singapore

Reviewed by:
Christian Bokhove, University of Southampton, United Kingdom
Shudong Zhang, Beijing Normal University, China

***Correspondence:**
Shima Salehi
salehi@stanford.edu

Specialty section:
This article was submitted to STEM Education, a section of the journal Frontiers in Education

Received: *07 May 2019*
Accepted: *13 September 2019*
Published: *27 September 2019*

Citation:
Salehi S, Cotner S, Azarin SM, Carlson EE, Driessen M, Ferry VE, Harcombe W, McGaugh S, Wassenberg D, Yonas A and Ballen CJ (2019) Gender Performance Gaps Across Different Assessment Methods and the Underlying Mechanisms: The Case of Incoming Preparation and Test Anxiety. Front. Educ. 4:107. doi: 10.3389/feduc.2019.00107

A persistent "gender penalty" in exam performance disproportionately impacts women in large introductory science courses, where exam grades generally account for the majority of the students' assessment of learning. Previous work in introductory biology demonstrates that some social psychological factors may underlie these gender penalties, including test anxiety and interest in course content. In this paper, we examine the extent that gender predicts performance *across* disciplines, and investigate social psychological factors that mediate performance. We also examine whether a gender penalty persists beyond introductory courses, and can be observed in more advanced upper division science courses. We ran analyses (1) across two colleges at a single institution: the College of Biological Sciences and the College of Science and Engineering (i.e., physics, chemistry, materials science, math); and (2) across introductory lower division courses and advanced upper division courses, or those that require a prerequisite. We affirm that exams have disparate impacts based on student gender at the introductory level, with female students underperforming relative to male students. We did not observe these exam gender penalties in upper division courses, suggesting that women are either being "weeded out" at the introductory level, or "warming to" timed examinations. Additionally, results from mediation analyses show that *across* disciplines and divisions, for women only, test anxiety negatively influences exam performance.

Keywords: gender, STEM equity, high stakes assessment, test anxiety, mediation analysis

INTRODUCTION

To effectively promote student groups who have been historically underrepresented in science, technology, engineering, and math (STEM), we need to provide students from different backgrounds equal opportunities to perform in these fields. Results from previous studies, however, demonstrate that schools are still unable to provide all students with equal opportunities, as

evidenced by gaps in performance based on gender and other demographic descriptors of student identities (McGrath and Braunstein, 1997; Kao and Thompson, 2003; DeBerard et al., 2004; Ballen and Mason, 2017). Demographic gaps in performance in higher education can be partly explained by demographic gaps in student incoming preparation (Sun, 2017; Salehi et al., 2019). However, they can also be due to biased education structures such as methods used to assess student performance in STEM fields (Stanger-Hall, 2012), introductory gateway courses that "weed out" students (Mervis, 2010, 2011), traditional uninterrupted lectures rather than high-structure active learning methods (Haak et al., 2011; Ballen et al., 2017b), feelings of exclusion (Hurtado and Ruiz, 2012), stereotype threat (Steele, 1997; Cohen et al., 2006), and discrimination (Milkman et al., 2015). While many demographics and identities remain underrepresented in STEM, such as certain racial and ethnic groups, and first-generation college students, the work described herein focuses broadly on women in STEM.

Previous work has demonstrated that using high stakes exams as an assessment method has disparate impacts on male and female students. Even after controlling for student incoming preparation, this work shows female students underperformed on exams across multiple introductory biology courses, due in part to test anxiety (Ballen et al., 2017a). This negative effect of anxiety on performance was observed only for female students, and only on exam assessments. Anxiety did not impact male student performance or female student performance on non-exam assessments such as homework and in-class assignments.

Women's underperformance on exams is troubling for two reasons in particular. First, exam scores usually constitute a high proportion of grades in introductory courses (Koester et al., 2016). If the primary assessment method in entry-level STEM courses leads to a "gender penalty" for female students, then institutions are creating an early obstacle that may prevent women from advancing to the upper-level subject material. Second, studies across STEM courses show that in some disciplines, low performance in introductory courses is disproportionately impactful for women, often resulting in the abandonment of their major, while men with similar performance are more likely to continue in the discipline (Grandy, 1994; McCullough, 2004; Rask and Tiefenthaler, 2008; Rauschenberger and Sweeder, 2010; Creech and Sweeder, 2012; Eddy and Brownell, 2016; Koester et al., 2016; Matz et al., 2017). Among women who perform well in introductory courses, Marshman et al. (2018) showed those who received high scores on a physics conceptual survey (or who were receiving A's) reported similar self-efficacy measures as male students with medium or low scores on the physics conceptual surveys (or who were receiving B's and C's). Therefore, female underperformance on exams, if generalizable across disciplines and over time, leads to a consequential gender performance gap that systematically disadvantages female students during their undergraduate pursuit of a degree.

Our previous work showed that women in introductory biology classes underperformed relative to men on exams, and that exam anxiety and interest in course content mediated the relationship between incoming preparation and exam performance (Ballen et al., 2017a). Until now, it was unclear whether the patterns we observed in undergraduate biology persist (1) in other disciplines, and (2) among students who have advanced beyond introductory science courses. First, biological sciences are among the most female-dominant fields in undergraduate STEM; ∼60% of undergraduate students in the life sciences are women (Neugebauer, 2006). If the gender gap in exam performance in introductory biology is due in part to the impact of test anxiety, this gap might be even more pronounced in male-dominated STEM fields where women are susceptible to negative social and learning experiences (e.g., tokenism, gender stereotypes about science abilities; Kanter, 1977; Miller et al., 2015).

Second, gender gaps in exam performance can be moderated by characteristics of the learning environment. Examples of characteristics that have documented impacts on student performance or experiences include group composition (e.g., gender ratio; Dahlerup, 1988; Sullivan et al., 2018), instructor traits (e.g., instructor gender: Crombie et al., 2003; Cotner et al., 2011, or attitude: Alsharif and Qi, 2014; Cooper et al., 2018b), and class size (Ballen et al., 2018). These characteristics may vary across disciplines, divisions, or even over a single semester. For example, upper division courses differ from lower division courses in a number of ways, and performance gaps present at the introductory level might not be apparent in more advanced courses. In upper division courses, student grades are less reliant on scalable multiple-choice exams; instead, the reliance on "lower stakes" assessments might ameliorate the negative impact of test anxiety on performance. Alternatively, or additionally, capable but test-anxious women may be weeded out at the lower division, or become acclimated to high stakes exams—or develop tools to counter test anxiety–as they progress through higher education.

In this study, we examined the generalizability of the exam gender gap across different STEM fields, and across both lower and upper division courses. We also studied how underlying social psychological mechanisms that have been previously studied in the context of student performance in STEM courses (e.g., test anxiety, interest in the course material Ballen et al., 2017a) change over time, and how they function as mediators of the gender gap in exam performance.

We address two multi-part questions as they apply *across different fields* (the College of Biological Sciences and the College of Science and Engineering) and *divisions* (lower and upper division courses):

1. ***Gender gap in different assessment methods (RQ1):*** (A) Do we observe a gender gap in performance across different assessment methods (i.e., exam, non-exam, laboratory, and course grades)? (B) To what extent can these potential gender gaps be explained by incoming preparation (as measured by students' American College Testing entrance exam score, hereafter ACT)?
2. ***Social psychological mediators of exam gender gap (RQ2):*** (A) How do test anxiety and interest in course content mediate performance outcomes on exams? (B) How do these two social psychological factors vary based on gender and over the course of a semester?

METHODS

Data Collection

The study is based on a secondary analysis of previously collected data that were provided **by CB and SC.** The IRB of the University of Minnesota exempted this study from the ethics review process (University of Minnesota IRB 00000800).

Class Performance

Administrative data were obtained from 5,864 students between 2015 to 2017. Courses included those offered by the College of Biological Sciences (CBS) or the College of Science and Engineering (CSE). A subset of the CBS data were explored in prior reports (e.g., Ballen et al., 2017a; Cotner and Ballen, 2017). CBS is a relatively small college (∼2,500 undergraduates) with a large percentage of women (the 2018 first year class was 66% female) and is restricted to biological fields including neuroscience, ecology, and genetics ("College of Biological Sciences," 2018). However, the lower-division courses involved in this study primarily target non-biology majors and only one of the courses enrolls students interested in pursuing biology as a major. These introductory biology courses not only include the standard curriculum, but also include courses that are customized to student interests such as "Environmental Biology: Science and Solutions," and the "Evolution and Biology of Sex," all of which fulfill introductory biology requirements for the university. The upper-division courses in CBS enroll predominantly students majoring in biology. CSE is a larger college (∼5,500 undergraduates) with a relatively small percentage of women (the 2018 percent of graduate and undergraduate females was 27.4% female; an all-time high percentage), and houses the departments of chemistry, physics and astronomy, chemical engineering and materials science, computer science and engineering, and the school of mathematics ("CSE: By numbers," 2018). Lower division courses included a mix of majors and non-majors, and upper division courses primarily served students who intended to major in the discipline (e.g., chemistry, computer science). We only included students who reported their gender in our sample ($N = 5766$) (**Table 1**).

TABLE 1 | Descriptive statistics.

	College of Biology Sciences (CBS)	College of Sciences and Engineering (CSE)
Lower	$N = 2,860$ URM: 277 (10%) Female: 1,520 (53%) First Generation: 417 (15%) Courses: Introductory Biology	$N = 2,409$ URM:258 (11%) Female: 1,065 (44%) First Generation: 384 (16%) Courses: General Chemistry, Computer Science, Math, and Physics
Upper	$N = 190$ URM: 19 (10%) Female: 113 (59%) First Generation: 20 (11%) Courses: Zoology, Evolution	$N = 307$ URM: 23 (7%) Female: 95 (31%) First Generation: 20 (7%) Courses: Advanced Chemical Engineering, Materials Science

We collected data from lower and upper division courses within the College of Biological Sciences (CBS) and the College of Science and Engineering (CSE). While lower division courses serve a variety of majors, upper division courses primarily enroll those in the respective college. URM, underrepresented minority students; First Generation, students who represent the first generation in their family to attend university.

In this sample, we compared (1) average exam scores, or scores on all high-stakes assessments that accounted for a relatively large portion of a student's grade, (2) average non-exam scores including in-class assignments, credit for participation, and group work (note: these scores do not include out-of-class homework), (3) average laboratory scores, where applicable, and (4) final course grades (i.e., student cumulative performance in the course based on their performance on all exam, lecture, and laboratory activities). For each of these items, we transformed all raw percentage scores into class Z-scores (a measure of how many standard deviations a value is from the class section's mean score) for ease of interpretation. We calculated Z-scores using the formula $Z\text{-score} = (X–\mu)/\sigma$, where X is the grade of interest, μ is the class mean score, and σ is the standard deviation.

Social Psychological Factors

In addition to performance data, for a subsample, we also examined change in exam anxiety and interest in course content over time. We conducted a survey at the beginning of the semester (pre-survey) and at the end of the semester before the final exam (post-survey). The survey included measures of student interest in course content as well as test anxiety (**Table 2**). For both metrics, we used multi-item constructs from Pintrich's et al. (1993) Motivated Strategies for Learning Questionnaire (MSLQ; **Table 2**). The MSLQ is a common tool for assessing motivated strategies for learning, with historically high reliability and validity across different student populations (e.g., Pintrich et al., 1993; McClendon, 1996; Büyüköztürk et al., 2004; Feiz and Hooman, 2013; Jakešová and Hrbáčková, 2014). Items on each subscale were rated on a 7-point scale (1 = not at all true for me to 7 = very true for me). Factor loadings of items were between 0.64 to 0.87 for interest in course content, and 0.73 to 0.89 for test anxiety. In the reliability study, the internal consistency alpha coefficient was calculated as 0.89 and 0.88, respectively, for these two subscales.

Statistical Analyses

For our analyses, we parsed our data across colleges and divisions: lower division CBS, lower division CSE, upper division CBS, and upper division CSE. We divided data across colleges because the students may be systematically different in each college in ways that impact our outcome variables. Differences across colleges in our sample are discussed in the data collection section. We also divided data across divisions for each college, as there may be a selection bias among those who pursue upper division courses. Upper division courses target students who have already chosen a STEM field for their major, while lower division courses also target non-major students. For each of the four sub-samples, we examined: **(RQ1.A)** the gender performance gap across different assessment methods (e.g., exam, non-exam, laboratory, and overall course grade); **(RQ1.B)** the impact of

TABLE 2 | Items used in a survey of students in courses offered by the College of Biological Sciences (CBS) and the College of Science and Engineering (CSE) at the University of Minnesota ($N = 3{,}368$).

Interest in science course content factor (alpha = 0.89)
It is important for me to learn what is being taught in this course.
I like what I am learning in this course.
I think I will be able to use what I learn in this course in later studies.
I think what I am learning in this course is useful for me to know.
I think that what we are learning in this course is interesting.
Understanding this subject is important to me.
Test anxiety factor (alpha = 0.88)
I am so nervous during a test that I cannot remember facts that I have learned.
I have an uneasy, upset feeling when I take a test.
I worry a great deal about tests.
When I take a test, I think about how poorly I am doing.

incoming preparation on assessment measures for men and women; **(RQ2.A)** the mediation effects of test anxiety and interest in course content on exam performance across genders; and **(RQ2.B)** how social psychological factors vary across genders and over time.

RQ1.A. Gender Gaps in Performance Across Different Assessment Methods

First, we analyzed *gender performance gaps* for different assessment methods without controlling for student incoming preparation and other demographic factors. These raw, "transcriptable" performance measures are what students see on their transcripts, use to assess their performance relative to their peers, and submit in graduate school applications. In order to examine the *gender gap* in performance, we used mixed-model regression analysis to predict student performance by gender without controlling for student incoming preparation. In this analysis, we included the fixed effect of gender, and the random effects of courses and sections to reflect the nested structure of the data (i.e., when sections are nested within courses).

RQ1.B. The Impact of Incoming Preparation on Gender Performance Gaps

Second, we examined the *gender gap* in student performance while controlling for student incoming preparation as well as their underrepresented minority (URM) status and first generation status (FGEN). Here we define URM students as those who are African American, Latino/a, Pacific Islander, and Native American. Incoming preparation was measured as students' American College Testing (ACT) score. The ACT is a standardized test that covers English, mathematics, reading, and science reasoning, and is commonly used for college admissions as well as in education research as a general measure of "incoming academic preparation." High schools in the United States vary substantially with respect to coursework, institution type (e.g., public, private, home-schooled), size, and grading scale. Admissions officers in higher education use tests such as the ACT to place student metrics such as grades and class rank in a national perspective (https://www.act.org). However, the location of public schools in the United States also dictates financial resources committed to them, such that a district with higher socio-economic status has more educational resources going to each individual student (Parrish et al., 1995). Thus, variation observed in ACT score can also be explained by socio-economic status of students or proxies thereof (e.g., minority status, first-generation status; Carnevale and Rose, 2013).

For this analysis, to find the simplest best-fitting model, we first started with a basic additive model. Then, we added different interaction terms between variables to this basic model, and tested whether addition of any interaction term would significantly improve the fit of the model. Our final model included gender (a factor with two levels: male = 0, female = 1), URM status (a factor with two levels: non-URM = 0, URM = 1), first generation (FGEN) status (i.e., whether the student was among the first generation in their family to attend university; a factor with two levels: continuing generation = 0, first generation = 1), and ACT score, as well as any interaction terms between these variables that improved the model fit significantly. Similar to the previous analyses, we also included the random effects of courses and sections.

RQ2.A. The Mediation Effect of Social Psychological Factors on Student Performance

For a subsample of students from whom we collected surveys, we used structural equation modeling (SEM) with lavaan R package (Rosseel, 2012) in order to test the structural relationship between incoming preparation, self-reported test anxiety, interest in course content, and exam performance for different genders. SEM is a statistical tool that allows us to address mechanisms underlying documented trends (Taris, 2002; Jeon, 2015). We used CFI, RMSEA, and SRMR to evaluate model fits. In this analysis, we normalized ACT score, test anxiety, and interest in course content for the whole sample. The normalized scores represent a measure of how many standard deviation a value is from the sample mean score. For students' general levels of test anxiety and course interest, we used data from the survey administrated at the beginning of the semester. The descriptive statistics of the subsample used in SEM are reported in **Table 3**.

RQ2.B. The Variation of Social Psychological Factors Across Genders and Time

To examine the variation of social psychological factors across genders and time, we analyzed how test anxiety and interest in course content vary over the semester for men and women. We used mixed-model multivariable regression analyses to regress either of these two psychological factors on gender and time points (beginning and end of the semester), while including the random effect of students.

RESULTS

RQ1.A. Gender Gaps in Performance Across Different Assessment Methods

Figure 1 shows the average normalized score for different assessment methods across genders. In the next section, we report

the sizes of gender gaps, and their significance across colleges and divisions for different assessment methods based on mixed-model single variable regression.

Consistent with the pattern observed in Ballen et al. (2017b), in lower division courses in CBS, women underperformed by a relatively small but significant margin on exams ($p = 0.033$) (**Table 4**). However, they significantly overperformed relative to men in non-exam ($p < 0.0001$) and laboratory measures ($p < 0.0001$). Due to their overperformance in these measures, women overperformed relative to men in overall course grades ($p = 0.044$). For CSE lower division courses, which include more male-stereotyped STEM disciplines such as physics, math, and chemistry, we observed the same trend of female underperformance on exams ($p < 0.0001$); of note, the size of the exam gender gap in CSE was three times that of CBS (−0.24 compared to −0.08 standard deviation). However, there was no gender gap in the non-exam measure ($p = 0.233$), and women significantly overperformed relative to men in the laboratory measure ($p = 0.033$). Due in part to the difference in the size of the gender gap in exams, as well as the differential weighting of exams in the overall course grade (e.g., Cotner and Ballen, 2017), women underperformed relative to men in overall course grades in lower division CSE ($p = 0.002$).

For upper division students in both CBS and CSE, we observed no influence of gender on exam performance ($p_{CBS} = 0.164$, $p_{CSE} = 0.987$, **Table 5**). However, in non-exam assessments, women marginally overperformed relative to men in CBS courses ($b_{CBS} = 0.28$, $p_{CBS} = 0.072$), and significantly overperformed in CSE couses ($b_{CSE} = 0.54$, $p_{CSE} < 0.0001$). On overall course grades, we did not observe gender differences across disciplines ($p_{CBS} = 0.108$; $p_{CSE} = 0.352$, **Table 5**). Due to a left skew in our data, to be conservative we also re-ran all the above analyses using non-parametric tests. The outcomes were the same, and results of non-parametric analyses were very similar to regression analyses reported here (**Tables S1, S2**).

TABLE 3 | Descriptive statistics of the data used for structural equation modeling.

	College of Biology Sciences (CBS)	College of Sciences and Engineering (CSE)
Lower	*N* = 1000 URM: 83 (8%) Female: 568 (57%) First Generation: 138 (14%) Courses: Introductory Biology	*N* =1871 URM: 205 (11%) Female: 788 (42%) First Generation: 273 (15%) Courses: General Chemistry, Computer Science, Math, and Physics
Upper	*N* =190 URM: 19 (10%) Female: 113 (59%) First Generation: 20 (11%) Courses: Zoology, Evolution	*N* = 307 URM: 23 (7%) Female: 95 (31%) First Generation: (7%) Courses: Advanced Chemical Engineering, Materials Science

We collected data from lower and upper division courses within the College of Biological Sciences (CBS) and the College of Science and Engineering (CSE) at University of Minnesota. We only included students in this analysis from whom we had complete data, including pre- and post-survey of social psychological factors.

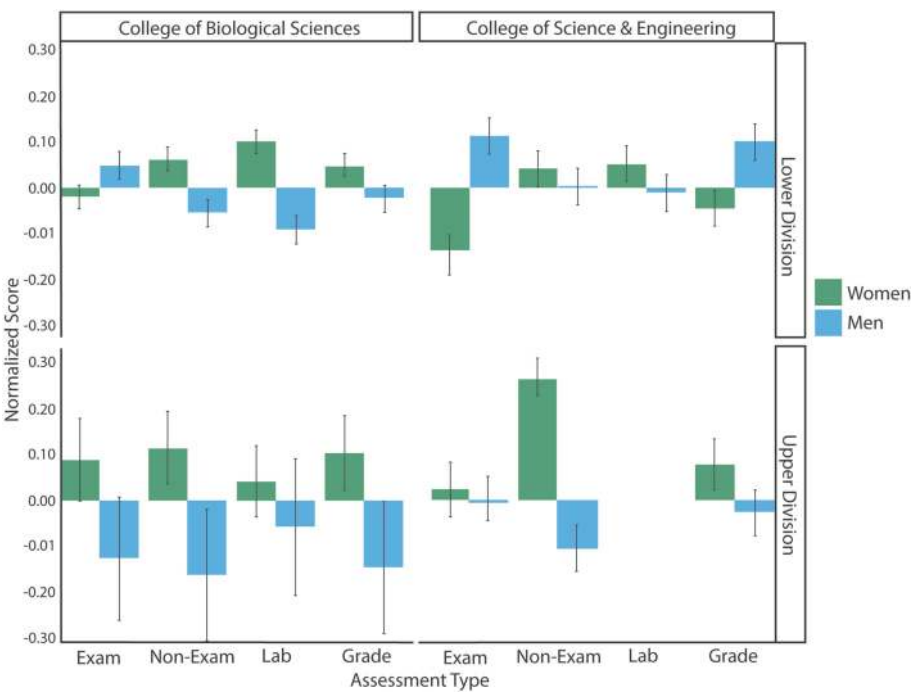

FIGURE 1 | Average normalized performance for women and men across different assessment methods (exam, non-exam assessments, laboratory, and overall course grade), disciplines (College of Biological Sciences, CBS; College of Science and Engineering, CSE), and divisions (lower division or upper division courses).

TABLE 4 | Regression estimates for the *gender gap* in performance in lower division courses in the College of Biological Sciences (CBS) and the College of Science and Engineering (CSE) across different assessment methods, including the overall course grade.

	CBS	CSE
Exam	b = −0.08 (0.04), $p = 0.033^{*}$	b = −0.24 (0.04), $p < 0.0001^{***}$
Non-exam	b = 0.16 (0.04), $p < 0.0001^{***}$	b = 0.06 (0.05), $p = 0.233$
Laboratory	b = 0.19 (0.04), $p < 0.0001^{***}$	b = 0.12 (0.05), $p = 0.033^{*}$
Grade	b = 0.07 (0.04), $p = 0.044^{*}$	b = −0.14 (0.04), $p = 0.002^{**}$

*A negative coefficient (b) indicates women underperformed relative to men in standard deviation; a positive b indicates women overperformed relative to men. This table reports raw values that are not adjusted for incoming preparation, and other demographic factors. Significant codes are: ***p < 0.001, **0.001 < p < 0.01, *0.01 < p < 0.05, †0.05 < p < 0.1.*

TABLE 5 | Regression estimates for the *gender gap* in performance in upper division courses in the College of Biological Sciences (CBS) and the College of Science and Engineering (CSE) across different assessment methods, including the overall course grade.

	CBS	CSE
Exam	b = 0.22 (0.16), $p = 0.164$	b = −0.002 (0.13), $p = 0.987$
Non-exam	b = 0.28 (0.15), $p = 0.072^{\dagger}$	b = 0.54 (0.12), $p < 0.0001^{***}$
Laboratory	b = 0.10 (0.16), $p = 0.520$	NA
Grade	b = 0.25 (0.16), $p = 0.108$	b = 0.12 (0.13), $p = 0.352$

*Note that none of the upper division courses in CSE had a laboratory component. Significant codes are: ***p < 0.001, ** 0.001 < p < 0.01, * 0.01 < p < 0.05, †0.05 < p < 0.1.*

RQ1.B. The Impact of Incoming Preparation on Gender Performance Gaps

To analyze what portion of gender performance gaps described in **Tables 4**, **5** are due to differences in incoming academic preparation, we used mixed-model multivariable linear regression with ACT as a measure of incoming preparation. In this analysis, we also controlled for URM and FGEN status of students. **Table 6** reports the coefficients of the simplest best fitting regression models predicting performance in each assessment method for lower division courses across colleges. In the following, we will focus on the effect of gender in this analysis. However, there are noteworthy effects of URM and first generation status on student performance that we discuss in detail in a forthcoming publication (Salehi et al., in preparation). Because our data violated the assumption of normality of residual distribution, we also analyzed the data using robust regression (Yaffee, 2002; Koller, 2015, 2016). The results of robust regression are reported in (**Tables S3, S4**). The results of these analyses were aligned with the following reported results.

For lower division courses, we found no significant gender gap in exam performance in CBS courses ($p = 0.404$) after controlling for incoming preparation (**Table 6**). However, even after controlling for incoming preparation, women performed 0.19 standard deviation lower than men on exams in lower division CSE courses ($p < 0.0001$). In contrast, women overperformed relative to men in non-exam and laboratory scores by 0.23 ($p < 0.0001$), and 0.22 standard deviation ($p < 0.0001$), respectively, in lower division CBS courses; and 0.17 standard deviation ($p = 0.004$) in laboratory scores in lower division CSE courses. Women also marginally overperformed by 0.10 standard deviation in non-exam scores of lower division CSE courses ($p = 0.090$). For the overall course grade, after controlling for incoming preparation, women significantly overperformed relative to men by 0.16 standard deviation in CBS courses ($p = 0.0001$), but they marginally underperformed by 0.09 standard deviation in CSE courses ($p = 0.057$).

TABLE 6 | The simplest best fitting models predicting performance in lower division courses across different assessment methods in the College of Biological Sciences (CBS) and the College of Science and Engineering (CSE).

		CBS	CSE
Lower	Exam	$b_{ACT} = 0.43$ (0.02), $p < 0.0001^{***}$ $b_{URM} = -0.07$ (0.06), $p = 0.180$ $b_{Gender} = 0.03$ (0.03), $p = 0.404$ $b_{FGEN} = -0.03$ (0.05), $p = 0.489$ $b_{ACTxGender} = 0.07$ (0.3), $p = 0.035^{*}$	$b_{ACT} = 0.47$ (0.02), $p < 0.0001^{***}$ $b_{URM} = -0.08$ (0.07), $p = 0.222$ $b_{Gender} = -0.19$ (0.04), $p < 0.0001^{***}$ $b_{FGEN} = -0.08$ (0.06), $p = 0.152$
	Non–Exam	$b_{ACT} = 0.09$ (0.03), $p = 0.003^{**}$ $b_{URM} = -0.14$ (0.06), $p = 0.038^{*}$ $b_{Gender} = 0.23$ (0.04), $p < 0.0001^{***}$ $b_{FGEN} = 0.004$ (0.06), $p = 0.94$ $b_{ACTxGender} = 0.10$ (0.04), $p = 0.018^{*}$	$b_{ACT} = 0.17$ (0.03), $p < 0.0001^{***}$ $b_{URM} = -0.20$ (0.10), $p = 0.045^{*}$ $b_{Gender} = 0.1$ (0.06), $p = 0.090^{\dagger}$ $b_{FGEN} = -0.02$ (0.08), $p = 0.814$
	Laboratory	$b_{ACT} = 0.13$ (0.02), $p < 0.0001^{***}$ $b_{URM} = -0.06$(0.07), $p = 0.419$ $b_{Gender} = 0.22$ (0.04), $p < 0.0001^{***}$ $b_{FGEN} = 0.02$ (0.07), $p = 0.773$ $b_{ACTxFGEN} = 0.15$ (0.06), $p = 0.023^{*}$	$b_{ACT} = 0.14$ (0.03), $p < 0.0001^{***}$ $b_{URM} = -0.23$ (0.1), $p = 0.019^{*}$ $b_{Gender} = 0.17$ (0.06), $p = 0.004^{**}$ $b_{FGEN} = -0.001$ (0.08), $p = 0.989$
	Grade	$b_{Act} = 0.30$ (0.03), $p < 0.0001^{***}$ $b_{URM} = -0.12$ (0.06), $p = 0.053^{\dagger}$ $b_{Gender} = 0.16$ (0.04), $p < 0.0001^{***}$ $b_{FGEN} = -0.03$ (0.05), $p = 0.617$ $b_{ACTxGender} = 0.08$ (0.04), $p = 0.027^{*}$	$b_{ACT} = 0.43$ (0.02), $p < 0.0001^{***}$ $b_{URM} = -0.10$ (0.07), $p = 0.150$ $b_{Gender} = -0.09$ (0.04), $p = 0.057^{\dagger}$ $b_{FGEN} = -0.10$ (0.06), $p = 0.074^{\dagger}$

*Each cell reports the simplest best fitting model. The simplest best fitting model includes only interaction terms if their addition improved the fit of the model significantly. For each predictor, we report the coefficient, the standard error of the coefficient in parentheses, and p-value of that coefficient. Positive coefficients for categorical variables of URM, FGEN, and gender indicate that URM students, FGEN students, and female students overperformed relative to their counterparts, and negative values mean they underperformed. Significant codes are: ***p < 0.001, **0.001 < p < 0.01, *0.01 < p < 0.05, †0.05 < p < 0.1.*

In upper division courses, after controlling for incoming preparation, we found no gender difference in exam performance for both colleges (**Table 7**). However, female students overperformed in non-exam measures significantly in upper division CSE courses, and marginally in CBS courses. They had on average 0.58 standard deviation higher non-exam score in upper division CSE courses ($p < 0.0001$), and 0.28 standard deviation in CBS courses ($p = 0.096$). Upper division CSE courses in this sample did not have lab components, and in CBS courses, we did not observe differences in lab scores ($p = 0.330$). For the overall course grade, there was no gender gap in CBS ($p = 0.178$), and marginal female overperformance of 0.21 standard deviation in CSE ($p = 0.095$). This marginal overperformance of females in CSE can be explained by their overperformance in non-exam assessments.

In summary, female students only underperformed on exams in lower division, introductory courses. After accounting for incoming preparation through ACT score, this gender gap in exam performance closed in one college (CBS), and decreased in size in the other (CSE). In other forms of assessment, if we observed any gender difference, it was female students outperforming their male counterparts.

TABLE 7 | Predictors of performance in upper division courses across different assessment methods in the College of Biological Sciences (CBS) and the College of Science and Engineering (CSE).

		CBS	CSE
Upper	Exam	$b_{ACT} = 0.31$ (0.08), $p = 0.0001^{***}$ $b_{URM} = -0.42$ (0.24), $p = 0.081^{\dagger}$ $b_{Gender} = 0.15$ (0.15), $p = 0.336$ $b_{FGEN} = -0.23$ (0.24), $p = 0.335$	$b_{ACT} = 0.52$ (0.09), $p < 0.0001^{***}$ $b_{URM} = -0.27$ (0.23), $p = 0.258$ $b_{Gender} = 0.09$ (0.12), $p = 0.477$ $b_{FGEN} = 0.0001$ (0.22), $p = 0.990$ $b_{ACTxURM} = -0.37$ (0.20), $p = 0.066^{\dagger}$
	Non-Exam	$b_{ACT} = 0.04$ (0.09), $p = 0.677$ $b_{URM} = -0.28$ (0.27), $p = 0.298$ $b_{Gender} = 0.28$ (0.17), $p = 0.096^{\dagger}$ $b_{FGEN} = -0.09$ (0.27), $p = 0.735$	$b_{ACT} = 0.11$ (0.08), $p = 0.21$ $b_{URM} = -0.43$ (0.24), $p = 0.067^{\dagger}$ $b_{Gender} = 0.58$ (0.13), $p < 0.0001^{***}$ $b_{FGEN} = 0.06$ (0.23), $p = 0.798$
	Laboratory	$b_{ACT} = 0.20$ (0.08), $p = 0.018^{*}$ $b_{URM} = -0.31$ (0.25), $p = 0.218$ $b_{Gender} = 0.15$ (0.16), $p = 0.330$ $b_{FGEN} = -0.057$ (0.25), $p = 0.823$	NA
	Grade	$b_{ACT} = 0.21$ (0.08), $p = 0.013^{*}$ $b_{URM} = -0.47$ (0.26), $p = 0.067^{\dagger}$ $b_{Gender} = 0.22$ (0.16), $p = 0.178$ $b_{FGEN} = -0.20$ 0.26), $p = 0.445$	$b_{ACT} = 0.49$ (0.09), $p < 0.0001^{***}$ $b_{URM} = -0.28$ (0.24), $p = 0.241$ $b_{Gender} = 0.21$ (0.13), $p = 0.095^{\dagger}$ $b_{FGEN} = 0.02$ (0.23), $p = 0.929$ $b_{ACTxURM} = -0.36$ (0.21), $p = 0.082^{\dagger}$

*Each cell reports the simplest best fitting model. The simplest best fitting model includes only interaction terms if their addition improved the fit of the model significantly. For each predictor, we have reported the coefficient, the standard error of the coefficient in parentheses, and the p-value of that coefficient. Positive coefficients for categorical variables URM, FGEN, and gender indicate that URM students, FGEN students, and female students overperformed relative to their counterparts, and negative values mean they underperformed. Significant codes are: ***$p < 0.001$, **$0.001 < p < 0.01$, *$0.01 < p < 0.05$, †$0.05 < p < 0.1$.*

RQ2.A. The Mediation Effect of Social Psychological Factors on Student Performance

Previous work demonstrates that test anxiety and interest in the course content exert gender-specific impacts on exam performance in introductory biology (Ballen et al., 2017b). To test whether these patterns persisted across different disciplines and divisions, we re-tested the same model on this larger sample using structural equation modeling (SEM) analysis. We fit the hypothesized model, shown in **Figure 2**, to four sub-samples of data (CBS lower division, CSE lower division, CBS upper division, CSE upper division), and used gender as a grouping variable to fit this model to the data of each gender separately.

We hypothesized that for women and men in each of these four sub-samples, exam performance is influenced by incoming preparation, text anxiety, and interest in course content. Furthermore, test anxiety and interest in course content are influenced by student incoming preparation. Therefore, this model suggests that incoming preparation influences student exam performance directly, as well as indirectly through test anxiety and interest in course content. In other words, test anxiety and interest in course content partially mediate the effect of incoming preparation on exam performance. By fitting this model to the data of each gender separately, we tested whether these mediation effects are different across genders for each sub-sample.

Acceptable ranges for SEM fit indices are: 0–0.07 for root mean square error (RMSEA), above 0.95 for comparative fit index (CFI), and 0–0.1 for standardized root mean square residual (SRMR) (Taris, 2002). This model had acceptable fit indices for all four subsamples, suggesting that it was an acceptable model to describe the variation in the data (CBS-lower: RMSEA = 0.064, CFI = 0.990, SRMR = 0.020; CBS-upper: RMSEA = 0.000, CFI = 1.00, SRMR = 0.017; CSE-lower: RMSEA = 0.057, CFI = 0.989, SRMR = 0.023; CSE-upper: RMSEA = 0.000, CFI = 1.000, SRMR = 0.021).

For women in CBS courses, test anxiety negatively influenced exam scores in both lower and upper divisions; for male students, however, test anxiety did not correlate with exam scores (**Figure 3**). Further, for CBS female students, ACT score was also negatively correlated with test anxiety. Therefore, this model suggests that incoming preparation influences student exam performance positively and directly, as well as indirectly through test anxiety (the red path in **Figure 3**). It is also notable that this indirect effect was stronger for upper division courses, as the negative relationship

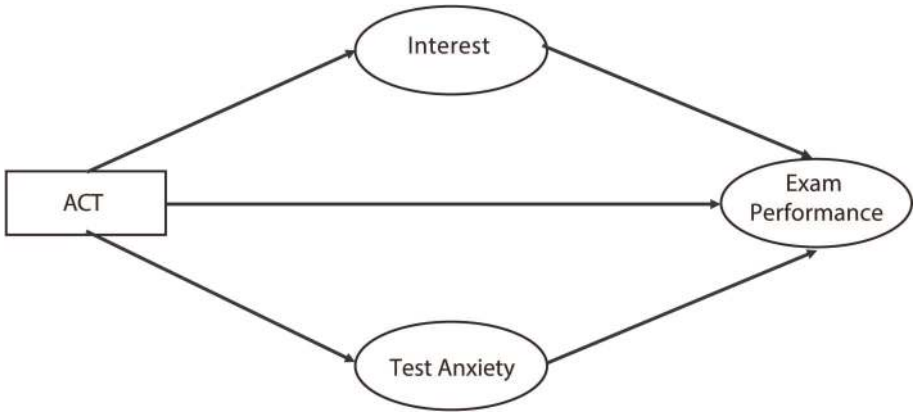

FIGURE 2 | Hypothesized model of the relationship among incoming preparation (ACT), social psychological factors (test anxiety, and interest in course content), and exam performance. We fit the structural equation model to data collected from men and women separately in the College of Biological Sciences (CBS) and the College of Science and Engineering (CSE), at both lower and upper division courses.

between test anxiety and female student exam score was stronger in upper division courses than in lower division courses. One standard deviation increase in test anxiety decreased exam score by 0.11 standard deviation in lower division courses and by 0.34 standard deviation in upper division courses.

Similarly, in CSE, test anxiety negatively correlated with exam scores for female students in both lower and upper divisions, but did not correlate with exam scores for male students in both divisions (**Figure 4**). However, unlike CBS, female student anxiety was correlated with their ACT scores only for the lower division courses, and not for the upper division courses. Therefore, the negative influence of anxiety is mediator for the indirect effect of ACT on exam for lower division CSE course. Like CBS, the negative influence of test anxiety on exam score increased in the upper division courses. One standard deviation increase in test anxiety decreased exam score by 0.14 standard deviation in lower division courses and 0.25 standard deviation in upper division courses. In summary, while the relationship between incoming preparation and test anxiety varied across women studying STEM at the University, we observed a consistently significant negative relationship between *test anxiety* and exam performance.

Our results suggest that regardless of discipline, exam performance for women was negatively influenced by their test anxiety, and surprisingly, this influence was *more* pronounced in upper division courses (**Figures 3**, **4**).

RQ2.B. The Variation of Social Psychological Factors Across Genders and Time

In all four sub-samples, except for CBS upper division courses ($p = 0.272$), women reported significantly higher levels of test anxiety than men (**Figure 5**). Women reported on average 0.35 standard deviation higher levels of test anxiety ($p = 0.0001$) than men in CBS lower division courses, and 0.6 standard deviation higher level of test anxiety in both lower and upper division CSE courses ($p < 0.0001$). Furthermore, in CSE upper division courses, test anxiety increased by 0.29 standard deviation over the semester ($p < 0.0001$).

Interest in course content was not a significant factor in predicting exam performance in any of the subsamples. That said, we still examined the variation in interest across genders and over the semester. We also found no gender difference in interest in upper division courses across both colleges ($p_{CBS} = 0.257$, $p_{CSE} = 0.665$), and no significant *change* in interest over the semester for CBS ($p_{CBS} = 0.900$, $p_{CSE} = 0.131$). However, in lower division courses, female students expressed 0.35 standard deviation higher interest in course content than male students in CBS courses ($p = 0.0001$), and 0.49 standard deviation lower interest in course content than male students in CSE courses ($p < 0.0001$). For all students, interest increased by 0.18 standard deviation in CBS lower division courses ($p = 0.023$), and decreased by 0.17 standard deviation in CSE lower division courses ($p = 0.005$). Changes over time in interest in lower division courses were not different between genders ($p_{CBS} = 0.648$, $p_{CSE} = 0.711$) (**Figure 6**).

DISCUSSION

Gaps in academic performance are attributable to a host of different external factors, including measures of academic preparation. However, even when accounting for preparation (e.g., via the ACT, SAT, or high-school grade-point average), achievement in some disciplines can be predicted by student characteristics such as gender, underrepresented minority status, and first generation status. We explored how factors *other than* these unidimensional categories of student identity—such as social psychological factors—impacted performance among students in science. We focused on mechanisms that underlie the gender-based performance gaps in different assessment methods across STEM fields and divisions.

We showed that women only underperformed in high stakes examinations in lower division introductory courses

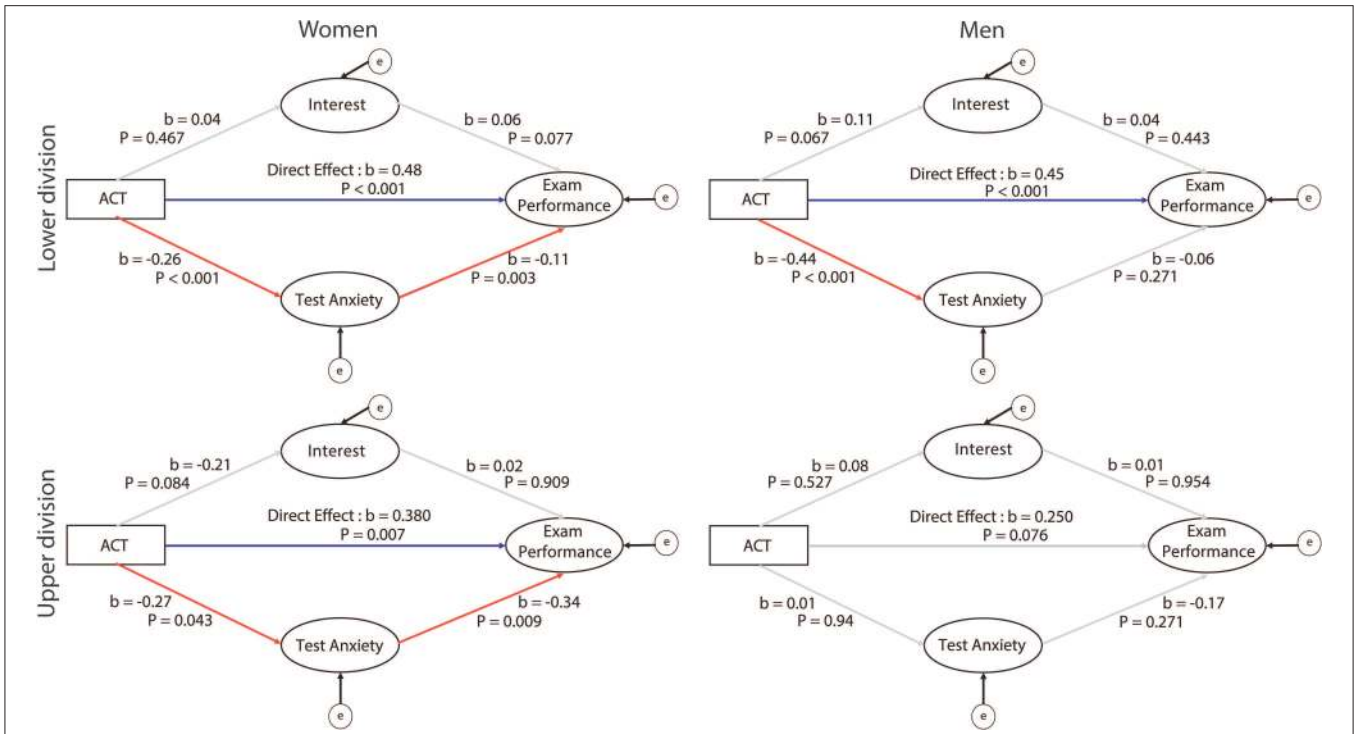

FIGURE 3 | Partial mediation analyses show differences in the significant effects of incoming preparation (ACT) on exam performance for female **(Left)** and male **(Right)** students in lower division and upper division courses in the College of Biological Sciences (CBS). Red arrows depict negative effects and blue arrows show positive effects. ACT has direct, positive effects on exam performance for all lower division students and for female upper division students. This effect was marginally significant for male upper division students. For female students in the lower and upper division courses, ACT negatively predicts test anxiety, which in turn influences exam performance. For male students at the lower and upper division, ACT negatively affects test anxiety, but test anxiety does not in turn influence exam performance. In the graphs, "e" circles indicate error terms in estimations of the structural equation model variables.

across multiple STEM fields. However, in non-exam and laboratory assessment methods in these introductory courses, either there was no gender gap or female students overperformed relative to their male peers. In CBS courses the gender gap in exam performance became non-significant when incoming preparation was accounted for. However, in CSE, even after controlling for incoming preparation, we observed a significant gender gap in exam scores. For upper division courses, unlike lower division courses, there was no gender gap in exam performance; and similar to lower division courses, in non-exam and laboratory assessment methods, either there was no gender gap, or female students overperformed relative to their male peers.

The gender difference in "transcriptable" grades in introductory courses in the two colleges could be due to several factors. First, the courses included in this study in CBS are some of a number of courses that meet the university's liberal education requirement for "biology-with-lab." The majority of the CBS courses included in this study do not serve as prerequisites for any other courses nor are they specifically required for most majors. All the CSE courses in this study are prerequisites for numerous courses and are required (or one of several challenging course options) for various majors. Therefore, the pressure to perform in the introductory level courses included in this study might be very different between the colleges. The grade pressure in a biology-with-lab course that is not a requirement for a student's major is likely lower than the grade pressure in courses that are considered gateways into a major. Further, this grade pressure may differentially impact the level of exam anxiety students feel. However, we did not see meaningful differences in test anxiety between the two colleges in these lower division courses.

We examined the mediation impact of test anxiety and interest in course content on gender performance on exams. The underperformance of women in lower division exams was explained in part by reported test anxiety. In upper division courses, which lack gender gaps in exam performance, test anxiety still negatively impacted exam performance for women, but not for men. For the remainder of this work, we further explore the phenomenon of anxiety—both general and test anxiety, especially as it pertains to gender-biased gaps in performance in STEM fields.

Test anxiety is common among university students; in one sample of undergraduates, 30.0% of males and 46.3% of females reported suffering from test anxiety. In this same report, students often declined seeking help from their peers or instructor for fear of the stigma associated with test anxiety (Gerwing et al., 2015). A meta-analysis of 126 studies found a negative correlation between test anxiety and performance, reporting that overall, students who reported low test anxiety overperformed relative

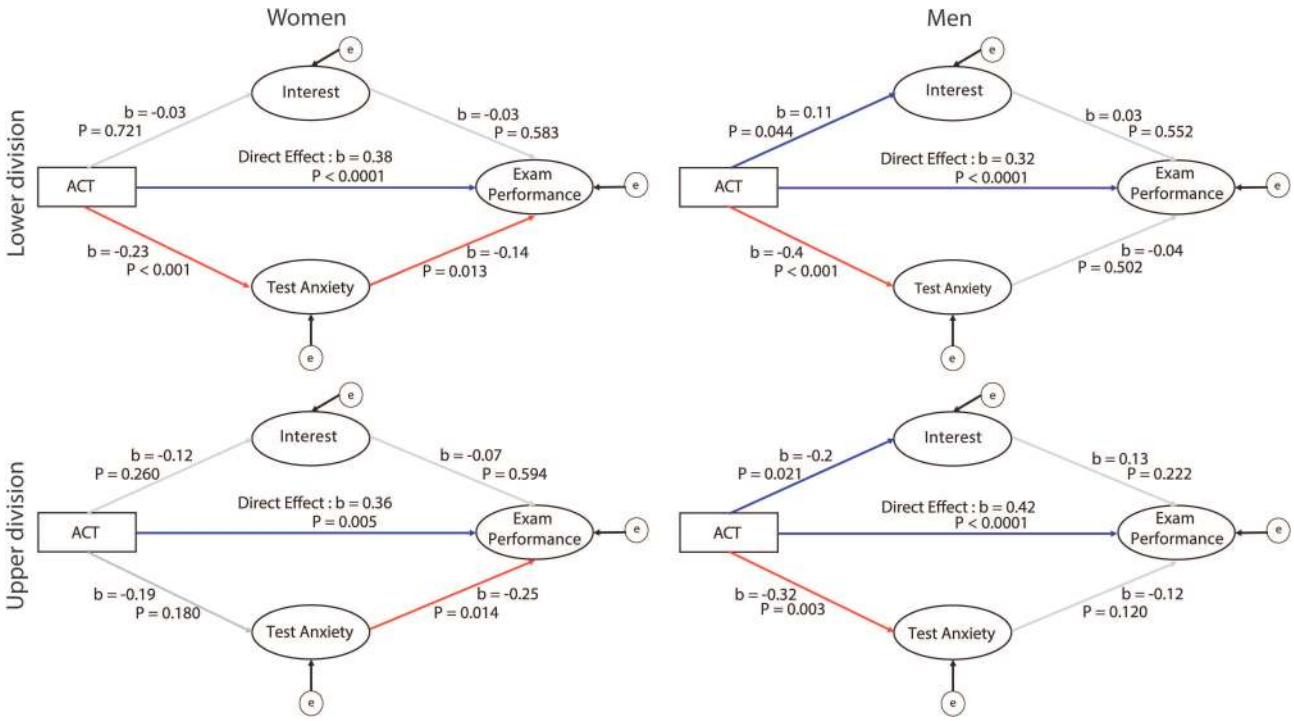

FIGURE 4 | Partial mediation analyses show differences in the significant effects of incoming preparation (ACT) on exam performance for female and male students in the College of Science and Engineering (CSE) in lower division and upper division courses. Red arrows depict negative effects and blue arrows show positive effects. ACT has direct, positive effects on exam performance for all lower division and upper division students. For female students at the lower division, ACT predicts test anxiety, which negatively predicts exam performance. For women in upper division courses, ACT does not predict test anxiety, but test anxiety negatively predicts exam performance. For male students at the lower and upper division, ACT negatively affects test anxiety, but test anxiety does not in turn influence exam performance. In the graphs, "e" circles indicate error terms in estimations of the structural equation model variables.

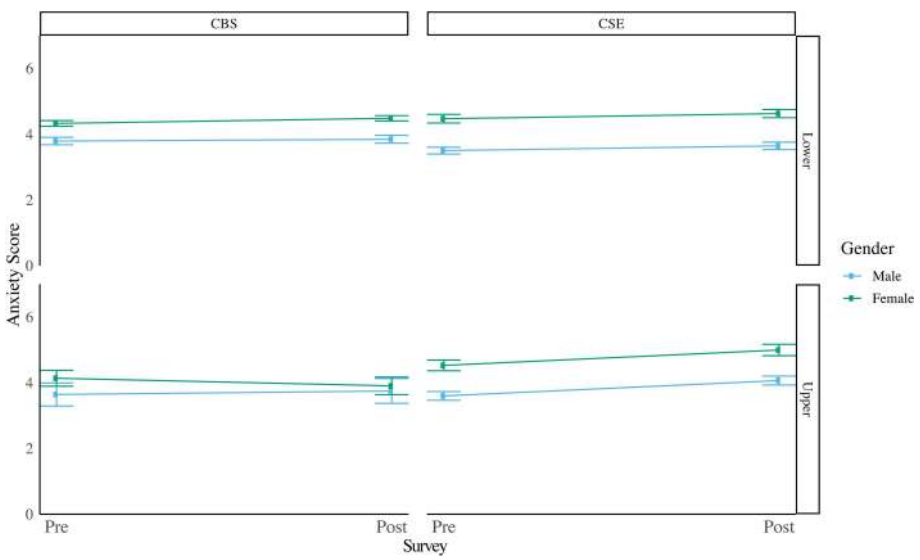

FIGURE 5 | Change in test anxiety over the course of the semester for students in CBS and CSE for women (green) and men (blue) in lower division courses **(Top)** and upper division courses **(Bottom)**. The survey was administered at the beginning of the semester (pre-survey) and at the end of the semester (post-survey; i.e., after students completed the last in-class test, but before their final exam). On average, women (green) reported higher levels of test anxiety than men (blue) in lower division courses in the College of Biological Sciences (CBS) and in both upper and lower divisions in the College of Science and Engineering (CSE). In upper division CSE, average test anxiety significantly increased for all students over the course of the semester.

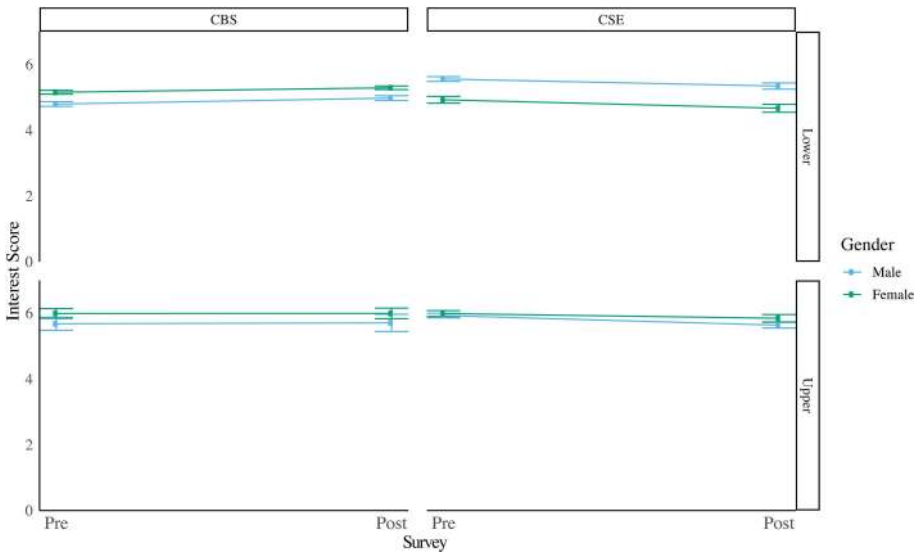

FIGURE 6 | Change in interest in course content over the course of the semester for students in CBS and CSE for women (green) and men (blue) in lower division courses **(Top)** and upper division courses **(Bottom)**. The survey was administered at the beginning (pre-survey) and at the end of semester (post-survey).

to students who reported high test anxiety by nearly half of a standard deviation (Seipp, 1991).

Further, women are more likely than men to be diagnosed with a generalized anxiety disorder (Wittchen et al., 1994; Kessler et al., 2005; Leach et al., 2008). Similarly, some investigators have documented higher levels of test anxiety in women than in their male peers (Osborne, 2001; Núñez-Peña et al., 2016; Ballen et al., 2017a). Our current work connects these threads by demonstrating that test anxiety negatively impacts exam performance for women, but not for men. Not only do these data confirm prior findings (Ballen et al., 2017a), but they elaborate on earlier work by identifying these trends in courses offered through multiple STEM disciplines besides biology.

While some hypothesize that heightened emotionality during an exam causes heightened anxiety, which in turn depresses performance (Maloney and Beilock, 2012; Ramirez et al., 2013), others suggest that it is the awareness of poor past performance that causes test anxiety (Hembree, 1990). In the first case, it is the anxiety that leads to the poor performance, and in the second, it is the poor performance that leads to the anxiety. Regardless of the origins of anxiety, there are certainly strategies instructors can use to minimize test anxiety and its impacts—strategies that are likely to benefit all students. And, given the demonstrated connection between introductory-level performance and retention in STEM (Seymour and Hewitt, 1997), it is worthwhile to pay closer attention to social psychological factors—such as test anxiety—that may disadvantage underrepresented groups.

How Can Instructors Address Student Anxiety?

It may be difficult to target each individual student's experience of anxiety, especially in the lower-division, high-enrollment courses. However, there are certain strategies instructors can employ to decrease the anxiety itself, or the impacts of the anxiety on a student's performance.

Rethinking assessment can be a helpful strategy that directly addresses test anxiety. Prior work in several introductory biology courses demonstrated that gendered performance gaps disappeared when exams were devalued in favor of the addition of multiple, lower-stakes assessments—possibly as a result of a reduction in test anxiety (Ballen et al., 2017a; Cotner and Ballen, 2017). The fact that women in our sample were more likely to underperform, relative to their male peers, on anxiety-inducing high stakes exams, combined with the fact that, across the board, women express higher levels of test anxiety, suggests that minimizing the impact of exams could lower performance gaps–such as those documented here.

Instructors can also create a classroom environment that minimizes general anxiety. Tanner (2013) discusses several instructional strategies for creating a welcoming classroom environment and reducing general class anxiety—from playing music before class to taking time to hear a range of student voices. Avoiding anxiety-inducing behaviors such as cold-calling on individual students (England et al., 2017; Cooper et al., 2018a), and opting for less stressful options such as calling on groups via randomly appointed spokespeople can minimize anxiety (Rocca, 2010). And creating a pattern of frequently encountered behaviors will allow students to adjust to the specific in-class expectations of the instructor (McCroskey, 2009). Finally, simply being transparent in expectations (about grading, test content, learning goals) can minimize anxiety (Neer, 1990). These are strategies that target student general anxiety in class, not particularly their test anxiety. While these two anxiety constructs can be positively correlated, they might differ significantly as well. Future works should explore whether and how reducing general

anxiety in class would impact test anxiety, and how this effect is moderated by demographic status.

Recommendations for Future Research

While there is compelling evidence that test anxiety, as well as anxiety in general, affects performance and retention, there is little if any work demonstrating the impacts of the above interventions on student anxiety, or the connections between lowered anxiety and improved performance. Thus, future work could measure the impacts of experimentally reducing anxiety on student outcomes such as performance, self-efficacy, sense of social belonging, and retention. For example, instructors could incorporate weekly quizzes, instead of or in addition to higher-stakes midterm exams, to test whether this reduces test anxiety, and in turn, improves performance for those impacted by test anxiety. Additionally, adding constructed response questions to summative assessments in large enrollment courses mitigates gender-biased performance outcomes (Stanger-Hall, 2012), and future work would benefit from exploring the impacts of different types of exam questions on student anxiety. Also, offering the option of retaking high stakes exams might reduce the anxiety associated with single metrics, as could extending the time allowed to complete exams.

With this current work, we cannot explain why the gendered gaps in performance disappear in upper division courses. Are women being "weeded out" after introductory courses, are they learning to cope over time, or benefitting from small classrooms (Ballen et al., 2018)? Also, because the populations are different—representing a greater range of majors in the introductory courses—we cannot rule out the possibility that the differences seen in lower division courses are driven largely by students not intending to major in science. These questions should be experimentally addressed, and will also benefit from longitudinal studies of individual students in the STEM pathway.

In this study we did not have any data about specific instructional practices employed in each particular course. Therefore, we could not examine how instructional practices in each course influenced gender gaps in different assessment methods. Second, the courses in our sample do not represent a cross-section of all courses offered in each college, nor were they selected to be contrasting cases of instructional practices. Our data collection was based on a convenience sample of instructors willing to share their data. Given that, we could only examine whether, on average, there existed gender gaps in different assessment methods in a set of different STEM courses in lower and upper divisions. Despite differences in student composition across two diverse colleges, the similar results we observed suggest these trends are generalizable to science majors and non-majors. Future studies might explore how different instructional practices affect demographic performance gaps, and which STEM fields have been more successful in employing these equitable instructional practices.

Finally, we used ACT as a measure of incoming preparation. We recognize that the ACT itself is a crude measure of student incoming preparation, and that it is also a high stakes examination. Other metrics, such as high-school ranking or GPA, might give a more accurate snapshot of a student's incoming preparation. Given the evidence from this study and previous studies, it is clear that the way in which we assess students should be reconsidered—not only within colleges and universities, but also in the admission process of higher education.

CONCLUSION

For investigators, there is still much work to be done to establish the salient connections between student affect, performance, and retention in STEM. And for instructors, it's clearly time to reconsider long-standing norms related to assessment strategies. Specifically, it may be time for a shift away from reliance on high stakes, timed examinations, which have negative effects on female students and may not be telling of a students' ability to succeed in a discipline. Rather, we encourage the use of evaluation that measures relevant skills, encourages growth, and allows instructors *and* students to better assess student progress.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

We received IRB exemption to work with student data from University of Minnesota, IRB 00000800.

AUTHOR CONTRIBUTIONS

SS, SC, and CB contributed to the conception and design of the study. CB organized the database. SS performed the statistical analysis. SS and CB wrote the first draft of the manuscript. SC wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was funded in part by a Research Coordination Network grant no. 1729935, awarded to SC and CB, from the National Science Foundation, RCN–UBE Incubator: Equity and Diversity in Undergraduate STEM.

ACKNOWLEDGMENTS

We thank Daniel Baltz for help with data organization, and the students who contributed performance data and their honest input on surveys.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2019.00107/full#supplementary-material

REFERENCES

Alsharif, N. Z., and Qi, Y. (2014). A three-year study of the impact of instructor attitude, enthusiasm, and teaching style on student learning in a medicinal chemistry course. *Am. J. Pharmac. Educ*. 78:132. doi: 10.5688/ajpe 787132

Ballen, C. J., Aguillon, S. M., Brunelli, R., Drake, A. G., Wassenberg, D., Weiss, S. L., et al. (2018). Do small classes in higher education reduce performance gaps in STEM?. *BioScience*. 68, 593–600. doi: 10.1093/biosci/biy056

Ballen, C. J., and Mason, N. A. (2017). Longitudinal analysis of a diversity support program in biology: a national call for further assessment. *Bioscience* 67, 367–373. doi: 10.1093/biosci/biw187

Ballen, C. J., Salehi, S., and Cotner, S. (2017a). Exams disadvantage women in introductory biology. *PLoS ONE* 12:e0186419. doi: 10.1371/journal.pone.0186419

Ballen, C. J., Wieman, C., Salehi, S., Searle, J. B., and Zamudio, K. R. (2017b). Enhancing diversity in undergraduate science: self-efficacy drives performance gains with active learning. *CBE-Life Sci. Educ*. 16:ar56. doi: 10.1187/cbe.16-12-0344

Büyüköztürk, S., Akgün, Ö. E., Özkahveci, Ö., and Demirel, F. (2004). The validity and reliability study of the Turkish version of the motivated strategies for learning questionnaire. *Educ. Sci. Theory Pract.* 14, 821–833. doi: 10.12738/estp.2014.3.1871

Carnevale, A. P., and Rose, S. (2013). *Socioeconomic Status*, *Race/Ethnicity, and Selective College Admissions*. Center on Education and the Workforce.

Cohen, G. L., Garcia, J., Apfel, N., and Master, A. (2006). Reducing the racial achievement gap: a social-psychological intervention. *Science* 313, 1307–1310. doi: 10.1126/science.1128317

Cooper, K. M., Downing, V. R., and Brownell, S. E. (2018a). The influence of active learning practices on student anxiety in large-enrollment college science classrooms. *Int. J. STEM Educ.* 5:23. doi: 10.1186/s40594-018-0123-6

Cooper, K. M., Hendrix, T., Stephens, M. D., Cala, J. M., Mahrer, K., Krieg, A., et al. (2018b). To be funny or not to be funny: gender differences in student perceptions of instructor humor in college science courses. *PLoS ONE* 13:e0201258. doi: 10.1371/journal.pone.0201258

Cotner, S., Ballen, C., Brooks, D. C., and Moore, R. (2011). Instructor gender and student confidence in the sciences: a need for more role models. *J. College Sci. Teach*. 40, 96–101.

Cotner, S., and Ballen, C. J. (2017). Can mixed assessment methods make biology classes more equitable? *PLoS ONE* 12:e0189610. doi: 10.1371/journal.pone.0189610

Creech, L. R., and Sweeder, R. D. (2012). Analysis of student performance in large-enrollment life science courses. *CBE Life Sci. Educ.* 11, 386–391. doi: 10.1187/cbe.12-02-0019

Crombie, G., Pyke, S. W., Silverthorn, N., Jones, A., and Piccinin, S. (2003). Students perceptions of their classroom participation and instructor as a function of gender and context. *J. High. Educ.* 74, 51–76. doi: 10.1353/jhe.2003.0001

Dahlerup, D. (1988). From a small to a large minority: women in Scandinavian politics. *Scand. Pol. Stud.* 11, 275–298. doi: 10.1111/j.1467-9477.1988.tb00372.x

DeBerard, M. S., Spielmans, G. I., and Julka, D. L. (2004). Predictors of academic achievement and retention among college freshmen: a longitudinal study. *Coll. Stud. J.* 38, 66–81.

Eddy, S. L., and Brownell, S. E. (2016). Beneath the numbers: a review of gender disparities in undergraduate education across science, technology, engineering, and math disciplines. *Phys. Rev. Phys. Educ. Res.* 12:20106. doi: 10.1103/PhysRevPhysEducRes.12.020106

England, B. J., Brigati, J. R., and Schussler, E. E. (2017). Student anxiety in introductory biology classrooms: perceptions about active learning and persistence in the major, *PLoS ONE* 12:e0182506. doi: 10.1371/journal.pone.0182506

Feiz, P., and Hooman, H. A. (2013). Assessing the Motivated Strategies for Learning Questionnaire (MSLQ) in Iranian students: construct validity and reliability. *Proc. Soc. Behav. Sci.* 84, 1820–1825. doi: 10.1016/j.sbspro.2013.07.041

Gerwing, T. G., Rash, J. A., Allen Gerwing, A. M., Bramble, B., and Landine, J. (2015). Perceptions and incidence of test anxiety. *Can. J. Scholars. Teach. Learn.* 6:3. doi: 10.5206/cjsotl-rcacea.2015.3.3

Grandy, J. (1994). Gender and ethnic differences among science and engineering majors: experiences, achievements, and expectations. *ETS Res. Rep. Ser*. 1994, i−63. doi: 10.1002/j.2333-8504.1994.tb01603.x

Haak, D. C., HilleRisLambers, J., Pitre, E., and Freeman, S. (2011). Increased structure and active learning reduce the achievement gap in introductory biology. *Science* 332, 1213–1216. doi: 10.1126/science.1204820

Hembree, R. (1990). The nature, effects, and relief of mathematics anxiety. *J. Res. Math. Educ.* 21, 33–46. doi: 10.2307/749455

Hurtado, S., and Ruiz, A. (2012). The climate for underrepresented groups and diversity on campus. *Am. Acad. Polit. Soc. Sci.* 634, 190–206. doi: 10.1177/0002716210389702

Jakešová, J., and Hrbáčková, K. (2014). The Czech adaptation of motivated strategies for learning questionnaire (MSLQ). *Asian Soc. Sci.* 10, 72–78. doi: 10.5539/ass.v10n12p72

Jeon, J. (2015). The strengths and limitations of the statistical modeling of complex social phenomenon: focusing on SEM, path analysis, or multiple regression models. *Int. J. Soc. Behav. Educ. Econ. Bus. Ind. Eng.* 9, 1594–1602.

Kanter, R. M. (1977). Some effects of proportions on group life: Skewed sex ratios and responses to token women. *Am. J. Soc.* 82, 965–990. doi: 10.1086/226425

Kao, G., and Thompson, J. S. (2003). Racial and ethnic stratification in educational achievement and attainment. *Ann. Rev. Soc.* 29, 417–442. doi: 10.1146/annurev.soc.29.010202.100019

Kessler, R. C., Chiu, W. T., Demler, O., Merikangas, K. R., and Walters, E. E. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Arch. Gen. Psychiatry*. 62, 617–627. doi: 10.1001/archpsyc.62.6.617

Koester, B. P., Grom, G., and McKay, T. A. (2016). Patterns of gendered performance difference in introductory STEM courses. *arXiv arXiv:*1608.07565.

Koller, M. (2015). *robustlmm: Robust Linear Mixed Effects Models*. R package version 2.1. Available online at: http://CRAN.R-project.org/package=robustlmm (accessed August 8, 2019).

Koller, M. (2016). robustlmm: an R package for robust estimation of linear mixed-effects models. *J. Stat. Softw.* 75, 1–24. doi: 10.18637/jss.v075.i06

Leach, L. S., Christensen, H., Mackinnon, A. J., Windsor, T. D., and Butterworth, P. (2008). Gender differences in depression and anxiety across the adult lifespan: the role of psychosocial mediators. *Soc. Psychiatry Psychiat. Epidemiol*. 43, 983–998. doi: 10.1007/s00127-008-0388-z

Maloney, E. A., and Beilock, S. L. (2012). Math anxiety: who has it, why it develops, and how to guard against it. *Trends Cogn. Sci.* 16, 404–406. doi: 10.1016/j.tics.2012.06.008

Marshman, E. M., Kalender, Z. Y., Nokes-Malach, T., Schunn, C., and Singh, C. (2018). Female students with As have similar physics self-efficacy as male students with Cs in introductory courses: a cause for alarm? *Phys. Rev. Phys.Educ. Res.* 14:20123. doi: 10.1103/PhysRevPhysEducRes.14.020123

Matz, R. L., Koester, B. P., Fiorini, S., Grom, G., Shepard, L., Stangor, C. G., et al. (2017). Patterns of gendered performance differences in large introductory courses at five research universities. *AERA Open*. 3:2332858417743754. doi: 10.1177/2332858417743754

McClendon, R. C. (1996). Motivation and cognition of preservice teachers: *MSLQ. J. Instruc. Psychol.* 23:216.

McCroskey, J. C. (2009). Communication apprehension: what have we learned in the last four decades. *Hum. Commun.* 12, 157–171.

McCullough, L. (2004). Gender, context, and physics assessment. *J. Int. Womens Stud.* 5, 20–30.

McGrath, M., and Braunstein, A. (1997). The prediction of freshmen attrition: an examination of the importance of certain demographic, academic, financial and social factors. *Coll. Stud. J.* 31, 396–408.

Mervis, J. (2010). Better intro courses seen as key to reducing attrition of STEM majors. *Science* 330:306. doi: 10.1126/science.330.6002.306

Mervis, J. (2011). Weed-out courses hamper diversity. *Science* 334:1333. doi: 10.1126/science.334.6061.1333

Milkman, K. L., Akinola, M., and Chugh, D. (2015). What happens before? A field experiment exploring how pay and representation differentially shape bias on the pathway into organizations. *J. Appl. Psychol.* 100, 1678–1712. doi: 10.1037/apl0000022

Miller, D. I., Eagly, A. H., and Linn, M. C. (2015). Women's representation in science predicts national gender-science stereotypes: evidence from 66 nations. *J. Educ. Psychol.* 107, 631–644. doi: 10.1037/edu0000005

Neer, M. R. (1990). Reducing situational anxiety and avoidance behavior associated with classroom apprehension. *South. J. Commun.* 56, 49–61. doi: 10.1080/10417949009372815

Neugebauer, K. M. (2006). Keeping tabs on the women: life scientists in Europe. *PLoS Biol*. 4:e97. doi: 10.1371/journal.pbio.0040097

Núñez-Peña, M. I., Suárez-Pellicioni, M., and Bono, R. (2016). Gender differences in test anxiety and their impact on higher education students academic achievement. *Proc. Soc. Behav. Sci.* 228, 154–160. doi: 10.1016/j.sbspro.2016.07.023

Osborne, J. W. (2001). Testing stereotype threat: does anxiety explain race and sex differences in achievement? *Contemp. Educ. Psychol.* 26, 291–310. doi: 10.1006/ceps.2000.1052

Parrish, T. B., Matsumoto, C. S., and Fowler, W. (1995). *Disparities in Public School District Spending 1989–90.* Washington, DC: National Center for Education Statistics Research and Development Report.

Pintrich, P. R., Smith, D. A. F., Duncan, T., and Mckeachie, W. J. (1993). Reliability and predictive validity of the Motivated Strategies for Learning Questionnaire (MSLQ). *Educ. Psychol. Measure*. 53, 801–813. doi: 10.1177/0013164493053003024

Ramirez, G., Gunderson, E. A., Levine, S. C., and Beilock, S. L. (2013). Math anxiety, working memory, and math achievement in early elementary school. *J. Cogn. Dev*. 14, 187–202. doi: 10.1080/15248372.2012.664593

Rask, K., and Tiefenthaler, J. (2008). The role of grade sensitivity in explaining the gender imbalance in undergraduate economics. *Econ. Educ. Rev*. 27, 676–687. doi: 10.1016/j.econedurev.2007.09.010

Rauschenberger, M. M., and Sweeder, R. D. (2010). Gender performance differences in biochemistry. *Biochem. Mol. Biol. Educ.* 38, 380–384. doi: 10.1002/bmb.20448

Rocca, K. A. (2010). Student participation in the college classroom: an extended multidisciplinary literature review. *Commun. Educ.* 59, 185–213. doi: 10.1080/03634520903505936

Rosseel, Y. (2012). *Lavaan: An R Package for Structural Equation Modeling and More. Version 0.5–12 (BETA)*. Ghent: Ghent University.

Salehi, S., Burkholder, E., Lepage, G. P., Pollock, S., and Wieman, C. (2019). Demographic gaps or preparation gaps? the large impact of incoming preparation on performance of students in introductory physics. *Phys. Rev. Phys. Educ. Res.* 15:020114. doi: 10.1103/PhysRevPhysEducRes.15.020114

Seipp, B. (1991). Anxiety and academic performance: a meta-analysis of findings. *Anxiety Res.* 4, 27–41.

Seymour, E., and Hewitt, N. M. (1997). *Talking About Leaving*. Boulder, CO: Westview Press.

Stanger-Hall, K. F. (2012). Multiple-choice exams: an obstacle for higher-level thinking in introductory science classes. *CBE Life Sci. Educ.* 11, 294–306. doi: 10.1187/cbe.11-11-0100

Steele, C. M. (1997). A threat in the air. *Am. Psychol.* 52, 613–629. doi: 10.1037//0003-066X.52.6.613

Sullivan, L. L., Ballen, C. J., and Cotner, S. (2018). Small group gender ratios impact biology class performance and peer evaluations. *PLoS ONE* 13:e0195129. doi: 10.1371/journal.pone.0195129

Sun, L. (2017). *How high school records and ACT scores predict college graduation* (Master of Science). Utah State University, Logan, UT, United States.

Tanner, K. D. (2013). Structure matters: twenty-one teaching strategies to promote student engagement and cultivate classroom equity. *CBE Life Sci. Educ.* 12, 322–331. doi: 10.1187/cbe.13-06-0115

Taris, T. W. (2002). BM Byrne, structural equation modeling with AMOS: basic concepts, applications, and programming Mahwah NJ: Lawrence Erlbaum, 2001 0-8058-3322-6. *Eur. J. Work Org. Psychol.* 11, 243–246.

Wittchen, H.-U., Zhao, S., Kessler, R. C., and Eaton, W. W. (1994). DSM-III-R generalized anxiety disorder in the National Comorbidity Survey. *Arch. Gen. Psychiatry* 51, 355–364. doi: 10.1001/archpsyc.1994.03950050015002

Yaffee, R. A. (2002). *Robust Regression Analysis: Some Popular Statistical Package Options.* ITS Statistics, Social Science and Mapping Group, New York State University.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.