Gendered Performance Gaps in an Upper-Division Biology Course: Academic, Demographic, Environmental, and Affective Factors

Victoria S. Farrar, Bianca-Yesenia Cruz Aguayo, and Natalia Caporale*
Department of Neurobiology, Physiology and Behavior, University of California Davis, Davis, CA 95616

ABSTRACT

Despite the existent gender parity in undergraduate biology degree attainment, gendered differences in outcomes are prevalent in introductory biology courses. Less is known about whether these disparities persist at the upper-division level, after most attrition is assumed to have occurred. Here, we report the consistent presence of gender equity gaps across 35 offerings (10 years) of a large-enrollment upper-division biology course at a research-intensive public university. Multilevel modeling showed that women's grades were lower than men's, regardless of prior GPA. These gender gaps were present even when controlling for students' race/ethnicity, socioeconomic status, first-generation college-going status, international status, and transfer status. Class size, gender representation in the classroom, and instructor gender did not significantly relate to course grades. Student questionnaires in a subset of offerings indicated gendered differences in course anxiety, science identity, and science self-efficacy, which correlated with grade outcomes. These results suggest that women experience differential outcomes in upper-division biology, which may negatively influence their persistence in STEM fields postgraduation. Our findings suggest that gender disparities are a systemic problem throughout the undergraduate biology degree and underscore the need for further examination and transformation of upper-division courses to support all students, even at late stages of their degrees.

INTRODUCTION

Women remain underrepresented in the science, technology, engineering, and mathematics (STEM) workforce, despite decades of diversification efforts. This disparity is already present at the undergraduate level, where men continue to enroll and graduate in STEM majors at a higher rate than women (De Brey *et al.*, 2021). The main exception to this binary gender (hereafter, "gender")[1] disparity is the life sciences, where in 2019, 63% of bachelor's degrees and 50% of doctoral degree recipients were women (De Brey *et al.*, 2021). However, inequities in the life sciences become evident at the postgraduate level, with women being underrepresented in both postdoctoral (43%) and tenure-track professor positions (31%) (De Brey *et al.*, 2021). The disparity at the professorial level increases as one goes up the ladder rank, with women comprising only 26% of full professors in the discipline. This disconnect between graduation and workforce parity is not unique to life sciences in academia; in medicine, women outnumber men as medical school matriculants, yet make up only 36% of practicing physicians (AAMC, 2018, 2019).

[1] Most studies on gender representation and discrimination to date compare binary genders (i.e., cisgender men and women), and do not consider transgender, non-binary, gender queer, or gender nonconforming individuals. These studies, like ours, are often limited to quantitative data collected by the universities in a binary manner, rather than information on how participants self-identify. To be consistent with prior literature, we use the terms "gender" to mean a binary comparison between "men" and "women", acknowledging these terms do not necessarily represent subjects' actual gender identities.

Sarah L. Eddy, *Monitoring Editor*
Submitted Mar 14, 2023; Revised Sep 15, 2023; Accepted Sep 27, 2023
CBE Life Sci Educ December 1, 2023 22:ar52
DOI:10.1187/cbe.23-03-0041
*Address correspondence to: Natalia Caporale (ncaporale@ucdavis.edu).

The dwindling representation of women in the life sciences workforce as one advances in positions is likely the result of a complex process of gender discrimination and bias that starts at the K–12 levels (Kanny *et al.*, 2014; Kuchynka *et al.*, 2022) and continues throughout college (Kanny *et al.*, 2014), job searches (Eaton *et al.*, 2020; Friedmann and Efrat-Treister, 2023), publication acceptance and citations (Fox and Paine, 2019; Ross *et al.*, 2022), grant funding rates (Hechtman *et al.*, 2018; Chaudhary *et al.*, 2021), and more. Gender inequities during college may be particularly deleterious, as this is when students are developing their professional identity and academic self-concept (Eccles and Wigfield, 2020). Such inequities have been described in classroom participation (Eddy *et al.*, 2014; Aguillon *et al.*, 2020; Bailey *et al.*, 2020; Nadile *et al.*, 2021), evaluation or assessment of ability by peers (Grunspan *et al.*, 2016; Bloodhart *et al.*, 2020), and levels of anxiety (Misra and McKean, 2000; Ballen *et al.*, 2017; Cooper *et al.*, 2023). These inequitable experiences for women may lead to gender disparities such as those seen in assessment outcomes like exam performance (Eddy *et al.*, 2014; Wright *et al.*, 2016; Ballen *et al.*, 2017) and course grades (Matz *et al.*, 2017; Malespina and Singh, 2023). These experiences are not unique to cisgender women, as nonbinary and transgender students also experience bias and discrimination in the STEM classroom (Garvey and Rankin, 2015; Casper *et al.*, 2022). For example, students of queer genders recount experiences of exclusion and lack of safety for their identities in biology classrooms, reducing their sense of belonging and interest in STEM fields (Casper *et al.*, 2022). As evidence of this effect, transgender and gender nonconforming students are more likely to leave STEM majors than their cisgender peers (Maloy *et al.*, 2022), a pattern is also seen in cisgender women compared with men (Koch *et al.*, 2022).

Studies examining binary gender disparities in student outcomes in biology courses have focused mostly on large, introductory “weed-out” or “gateway” courses (e.g., Eddy *et al.*, 2014; Freeman *et al.*, 2014; Ballen *et al.*, 2017; Wilton *et al.*, 2019). This research underscores the importance of introductory coursework, as students’ early experiences and outcomes have impacts on their persistence in STEM (Seymour and Hewitt, 1997; Ost, 2010). These studies have focused on pedagogical (e.g., assessment types: Cotner and Ballen, 2017; active learning structures: Wilton *et al.*, 2019) and environmental factors (e.g., class size: Ballen *et al.*, 2018; Bailey *et al.*, 2020) within the classroom, as well as student-affective factors that may relate to the observed gendered equity gaps at the introductory biology level (Eddy and Brownell, 2016) (e.g., test anxiety: Ballen *et al.*, 2017; Harris *et al.*, 2019; Salehi *et al.*, 2019). In contrast to the abundance of research on these factors in lower-level courses, very little is known about equity gaps and the structural, institutional, course-level, and student-level factors that may be associated with them in upper-division courses.

An argument can be made that at many large institutions, upper-division biology courses, especially those required for graduation in the major, have characteristics that can lead to them serving as additional “gateway” or “weeder” courses, impacting retention of students. While it is often assumed that upper-division coursework entails small enrollment seminars or elective courses covering narrow topics, many required courses for life science majors can be large (>100 students per offering) and because of this, less personal, typically taught through lecture-only formats and assessed through high-stakes exams (see examples in Creech and Sweeder, 2012). Thus, these courses share many of the attributes of lower division STEM courses that have been shown to be particularly detrimental to women’s performance. First, perceived representation in the classroom and instructional team matters to students, as women tend to receive higher grades when classes are instructed by women, and when the class has a large percentage of women (Bailey *et al.*, 2020; Bowman *et al.*, 2022). Second, stereotype threat may affect women’s performance on exams as they grapple with negative messages about women’s aptitudes for or belonging in science (Appel *et al.*, 2011). Stereotype threat may be increased by grade penalties[2] faced earlier in the major and may compound as students advance to higher level courses. Women tend to receive lower grades in large STEM lecture courses relative to their non-STEM courses than men (Matz *et al.*, 2017; Witteveen and Attewell, 2020), which may lead to lowered feelings of scientific self-efficacy and discouragement from STEM majors. Third, assessment type can contribute to gender gaps, as women tend to perform better on low-stakes assignments than high-stakes exams with challenging question structures (Wright *et al.*, 2016; Ballen *et al.*, 2017; Cotner and Ballen, 2017) and report higher test anxiety than men (Ballen *et al.*, 2017; Salehi *et al.*, 2019). Lastly, large class size has been associated with lower grades for women (Ballen *et al.*, 2018; Odom *et al.*, 2021). Given these shared factors, gendered equity gaps may perpetuate beyond introductory biology into the upper-division coursework.

There are only a limited number of studies on gendered equity gaps in biology that include upper-division courses (Salehi *et al.*, 2019; Bailey *et al.*, 2020; Malespina and Singh, 2023) and even fewer that focus on upper-division biology specifically (Rauschenberger and Sweeder, 2010; Creech and Sweeder, 2012). Across these five studies, there is mixed evidence for gendered grade disparities. Using multilevel modeling, Rauschenberger and Sweeder (2010) found that women had lower predicted grades than men in biochemistry courses when controlling for academic factors such as prior GPA, prerequisite course enrollment, and major. Creech and Sweeder (2012) extended this work to multiple biology and chemistry courses and found similar trends; models predicted significantly lower grades for women in nine out of 12 courses examined. Other studies, however, did not uncover similar gendered equity gaps. In upper-division zoology and evolution courses, no gender disparities in exams or overall course grades were found, despite being present in introductory biology at that institution (Salehi *et al.*, 2019). Similarly, Malespina and Singh (2023) found no significant difference in grade anomaly (i.e., the difference between course grade and prior GPA) between men and women in the two upper-division biology courses that they examined, while they were present in several lower division courses. Salehi *et al.* (2019) offer a few possible explanations for

[2] Grade penalties refer to student grades that are lower than what one may expect given the student's overall performance in other courses (Koester *et al.*, 2016; Matz *et al.*, 2017). Specifically, grade anomalies are calculated as the student's grade in the course minus the average of their grades obtained in all their other courses taken on the same quarter/semester (GPAO) or their overall cumulative GPA. If the grade anomaly is negative, it is called a “grade penalty” if positive, it is called a “grade bonus”.

the gender parity; 1) perhaps women with lower performance or high test anxiety left the major before reaching upper-division coursework; 2) women may have gained coping strategies since their introductory courses; or 3) women may benefit from the smaller class sizes typical of many upper-division courses. Given these initial findings, more work characterizing gendered equity gaps in upper-division courses and the environmental factors and affective factors that may underlie them, if present, is needed. Furthermore, these studies need to be conducted at a larger variety of institutions, as all publications to date focused on primarily white institutions (Bourke, 2016).

Although less studied, student outcomes in upper-division courses can also have cascading impacts on graduation and postgraduate opportunities. While it is often assumed that most attrition has taken place by the time students have reached upper-division coursework, students still leave STEM majors at this stage. Data on attrition from STEM after the third year are scant, but some studies suggest that late-stage attrition rates can range from 10 to 30% across all majors (Aulck and West, 2017). At our institution, overall attrition rates past the third year ranged from 3 to 7% for incoming freshmen and from 8 to 15% for transfer students (AggieDash, 2022; period 2000-2016), a population which is especially affected by upper-division outcomes. Even if students remain in the institution, they may still switch out of STEM majors later in their degree. Such late-stage major shifts can have negative repercussions on time to graduation and graduating grade point average, which would impact postgraduate opportunities. For example: at one institution, students who changed majors in the third year or beyond tended to have lower graduation rates, took longer to graduate, and/or had lower grades at graduation (Foraker, 2012).

Any impacts on graduation outcomes can also affect students' competitiveness for, and interest in, STEM careers. This impact may be especially strong for women, as women who graduate with biomedical degrees have been shown to be more likely to have shifted their career plans to ones requiring fewer years of postgraduate study and less likely to aspire to attend medical school than men (Rosenzweig *et al.*, 2021). Taken together with the fact that women's persistence in the sciences may be more sensitive to grades than men's (Ost, 2010), gendered differences in upper-division course grades may significantly alter women' postgraduate career plans and aspirations. In the context of biological disciplines, women's experiences in the last years of their undergraduate biology degree may be contributing to the lacking gender representation in the life science workforce, despite the gender parity in bachelor's degree attainment.

Research Questions

In this study, we examined gender gaps in upper-division courses by conducting multilevel modeling and implementing student questionnaires in a large-enrollment upper-division biology course, focusing on the following four research questions:

1. To what extent are gendered equity gaps, like those seen in introductory biology courses, present in large upper-division biology courses?
2. How do academic factors, such as prior grade point average (GPA), major, and number of STEM units taken, relate to student outcomes and gendered equity gaps?
3. How are other demographic factors associated with student course outcomes and gendered equity gaps?
4. What environmental and affective factors may be associated with the observed gendered performance gaps?

METHODS

Course details

We analyzed student grades in 35 individual offerings of an upper-division human physiology course at a large, research-intensive, land-grant, minority-serving public university on the quarter system. One to two sections of this course are typically offered every fall, winter, and spring quarter, with this study comprising 27 quarters, from Fall 2010 to Spring 2019 (a 10-year period). This research was approved by IRB (approval # 1002498; University of California, Davis).

This large human physiology course broadly covers the nervous, cardiovascular, renal, respiratory, muscular, immune, endocrine, and reproductive systems. Prerequisites for this course include introductory biology and chemistry. Although taught by a variety of different instructors, the course follows a consistent structure across offerings: five 50-minute lectures per week with course grades being determined exclusively by two to three multiple-choice exams and a final exam. Enrollment in each course offering averaged 377 students, ranging from 177 to 528 students. Each offering of the course is typically co-taught by 2–3 instructors (61% of offerings had two instructors, 32% had three). Forty-nine percent of offerings had one or more women instructors, and of these, two offerings had an all-women instructional team. These offerings include five that were co-taught by one of the authors of this study (N.C.). Instructor sex data was accessed from the university registrar.

The course is required for most majors in the College of Biological Sciences and is also taken by prehealth students in other colleges in preparation for biomedical careers. The top six most populous majors across our dataset were: 1) Neurobiology, Physiology and Behavior (a Neuroscience-type major; 16% of total students), 2) Biological Sciences (15%), 3) Psychology (Biological Emphasis; 12%), 4) Biochemistry and Molecular Biology (8%), 5) Animal Science (a pre-veterinary medicine type major; 7%) and 6) Clinical Nutrition (6%). All these majors include human physiology as a "depth subject matter" requirement, except Animal Science, where students may elect to take a general vertebrate physiology course instead. These majors all require an organic chemistry and introductory physics series, except Clinical Nutrition and Animal Science, which do not require physics. Most students in our dataset were in their third (56%) or fourth year (35%) of college, with second-year students accounting for the remaining 9% of the students in our sample. The third- and four-year contingents included many community college transfer students (30 and 33%, respectively). While the original sample included freshmen and graduate students (accounting for 0.3% of the student sample), these were excluded from the analysis (see sample description in the next section).

Student demographic data

We accessed all course grades and demographic data from the university registrar. The total number of students in our initial sample for the Fall 2010 to Spring 2019 academic quarters (excluding summer) was 13,391. If students took the course more than once, we selected only the first term in which they took the course (i.e., we excluded any additional enrollments or repetitions of the course after a student's first enrollment; each student was only included once). We removed freshmen, graduate students, and students whose class standing was unknown ($n = 40$, 6, 5, respectively, accounting for a combined 0.038% of sample) as this is an upper-division course not traditionally available to first-year students and we were interested in the undergraduate population. As a result, our prefiltering sample size was 13,340. We then applied the following exclusion criteria (note that some students fell in more than one exclusion criteria): we excluded students who did not receive a standard letter grade in the course ($n = 104$), and those for whom we did not have any prior cumulative GPA data ($n = 49$). Prior GPA was defined as students' cumulative GPA at our institution ending in the term before the term in which they took our course (e.g., if a student took human physiology in spring quarter 2011, their prior GPA would be their cumulative GPA at the end of winter quarter 2011). Of students without prior GPA data, the largest percentage corresponded to transfer students in their first term at the four-year institution (i.e., they took upper-division physiology in the term they were admitted to the university; $n = 22$; 45% of all students filtered for missing prior GPA data). Lastly, we excluded students whose gender data was not available ($n = 3$), yielding a final sample size of 13,184 for the study.

Students who did not have data on race/ethnicity ($n = 248$; 1.89% of final sample), socioeconomic status ($n = 2{,}016$; 15.3% of final sample) or first-generation college student status ($n = 394$; 2.99% of final sample), were not excluded but instead their demographics were defaulted to the majority group for those variables (e.g., students for whom no ethnicity data was available were coded as white/non-PEER; no first-generation information were coded as continuing-generation/non–first-generation, and those with no socioeconomic information were coded as non–low socioeconomic status). Of those students affected, most were only missing one demographic variable (Supplemental Table S1). By coding missing demographic variables conservatively for these students, we were able to evaluate the impact of other facets of their identities without excluding these students completely. The significant main effects in the best fit model (see "*RQ2*" below) did not differ when these students were excluded completely, rather than being conservatively coded (Supplemental Table S2, compared with Table 5). Thus, we retained these students in the dataset to preserve sample size and evaluate the impacts of other identities these students may hold.

Of the 13,184 students included in this analysis, 66% were women (ranging from 59 to 74% across offerings; mean 66 ± 4%). The enrollment data available to us came from the university's registrar, which during the period of the study only collected gender data using a binary definition of sex ("male" versus "female"). Not having access to student self-reported gender identity, we decided to use the available binary data and employ the terms "women" and "men" throughout the text to be consistent with prior literature on binary gender gaps and because we are interested in the effects of student gendered experiences on course outcomes. We acknowledge that the limited binary classification of registrar data obscures the full spectrum of gender identities and does not accurately represent gender identity for all students. Thus, our findings may oversimplify the effects that diverse gender identities may have on academic performance (Cooper *et al.*, 2019) and may also miss factors that specifically lead to performance gaps for gender nonconforming students (Maloy *et al.*, 2022).

Within the conservatively coded dataset, 29% of the students were transfer students, 28% were English second language (ESL) students, 43% were first generation students, and 30% were low socioeconomic status students. In addition, 19% of the students in this dataset were categorized as persons excluded because of ethnicity or race (PEER; Asai, 2020), according to ethnicities provided and defined by the institutional registrar. The definition of PEER mirrors the definition of "underrepresented minorities" used by the National Science Foundation (2019) and included students from the following backgrounds: Black/African American, Indigenous/Native American Alaska Native, Latinx/Chicanx/Hispanic American, or a mix of these racial/ethnic identities. Demographic details by gender can be found in Table 1. Additional details on the exact representation of students by ethnicity and race can be found in Supplemental Table S3.

Statistical analysis

All descriptive statistics, longitudinal analysis, and linear modeling were performed in the R statistical language (R Core Team, 2020, v.4.2.1). All correlations (Pearson) were calculated using the "corrplot" package in R (Wei and Simko, 2021).

For all regression models, we assessed whether the residuals met the assumptions of normality and homoscedasticity, as is common practice (Zuur *et al.*, 2009). Given the properties of grades and categorical demographic data, none of our models met these assumptions. Thus, we implemented a robust estimation approach, which differentially weights residuals from outliers to address these deviations from assumptions and reduces the effect of outliers on the model. In R, this can be done using the "robustlmm" package (Koller, 2016).

When implementing model selection to identify the final models reported in the paper, it was not feasible to use robust models, as it is extremely complex and most common statistics software (including R), do not include a mechanism for this (Koller, 2016). Thus, all model selection processes were conducted using nonrobust linear models. Once a model was selected, we then estimated the final model fit and coefficients

TABLE 1. Student demographics disaggregated by gender

	Men	Women[a]
Sample size (*n*)	4443	8741
Transfer students (%)	34.1	25.9**
Persons excluded because of ethnicity or race (PEER) (%)[b]	18.4	19.5
English as a second language (ESL) (%)	27.1	28.9*
First-generation college students (%)	42.1	44.1*
Low socioeconomic status (%)	28.1	30.2*
International students (%)	4.4	4.8
Average overall prior GPA	3.09	3.07

[a]Significant χ^2 test by gender

[b]PEER students are defined as Black/African American, Latinx or Chicanx, American Indian/Indigenous or a mix of these ethnic identities. (Asai 2020).

* $p < 0.05$, ** $p < 0.01$

using robust regressions and reported those coefficients in the paper. (To illustrate the difference between the robust and nonrobust models, Supplemental Table S7 shows the coefficients calculated using robust and nonrobust methods for the two main models of the paper (Model I and Model II, see below), with the final robust weights being described in Supplemental Table S8).

RQ1: To What Extent are Gendered Equity Gaps, Like Those Seen in Introductory Biology Courses, Present in Large Upper-Division Biology Courses?

Initially, we examined overall trends in gendered differences in performance across the full dataset (n = 35 offerings, 13,184 students) using descriptive statistics and simple linear modeling. We compared overall average course grades across all offerings using two-tailed t tests, as well as compared the gendered performance gap in our course with that observed in our institution's introductory biology course for the subset of students who took introductory biology at our institution. We also plotted descriptive statistics of the distribution of final letter grades in upper-division human physiology, as well as explored gender performance gaps over time. Lastly, we created a simple multilevel model with course grade as the dependent variable and gender as the independent variable (*CourseGrade ~ Gender + (1|Offering)* + ε) to initially explore the impact of student gender on grade while controlling for the clustered nature of students in our data within course offerings (Theobald, 2018). These initial approaches established the presence and consistency of the gendered gaps in performance in this course.

RQ2: How do Academic Factors, such as Prior Grade Point Average (GPA), Major, and Number of STEM Units Taken, Relate to Student Outcomes and Gendered Equity Gaps?

Models aiming to examine factors that may impact student course grades often include a variable to control for prior academic preparation. Studies of lower division STEM courses often use incoming high school GPA and/or standardized test scores (e.g., Freeman *et al.*, 2014; Koester *et al.*, 2016; Matz *et al.*, 2017). When studying upper-division courses, prior academic performance at the university likely correlates more with student course outcomes as it better reflects students' course experiences at the institution. Indeed, even in introductory courses, Koester *et al.* (2016) found that a measure of cumulative university GPA correlated more strongly with course grades than did high school GPA or standardized test scores. Cumulative prior GPA has also been used as a control for academic preparation in other studies of student outcomes at the upper-division level (Creech and Sweeder, 2012; Salehi *et al.*, 2019). Furthermore, later in the analysis we used AIC to compare models that included prior GPA and models based on various combinations of introductory biology and chemistry grades and found that prior GPA yielded the best fit model. An example of such comparison is shown in Supplemental Table S4.

To initially explore the relationship between prior academic performance (via prior GPA) with gender and course grades, we extended the simple multilevel model from RQ1 above to include prior GPA as an independent variable (*CourseGrade ~ PriorGPA + Gender + (1|Offering)*). As we posited that student gender could affect the relationship between prior GPA and course outcomes, we also added an interaction between prior GPA and gender, resulting in the following model:

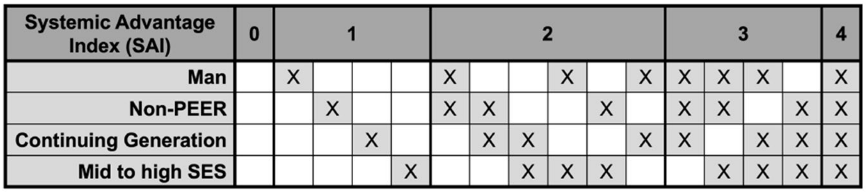
$$CourseGrade \sim PriorGPA + Gender + PriorGPA * Gender + (1 \mid Offering)$$

Our dataset included a large variety of majors (110), many of which differ in their introductory STEM prerequisites. This variance in STEM major requisites could lead to differences in prior GPAs that are reflective of both difference in preparation for upper division STEM courses and differences in grading practices across disciplines (Sabot and Wakeman-Linn, 1991; Ahn *et al.*, 2019). Thus, these major differences could explain any mismatch between prior GPA and course grades observed between men and women, especially if there were unequal distributions of men and women across the majors. To analyze this issue, we identified the six most populous/common majors in the course (which account for 62% of all students in our sample). Three of these majors belonged to the College of Biological Sciences, while the other three belonged to other colleges at the institution. Across the student sample, women were less likely than men to have majors in the College of Biological Sciences (52% of men vs. 39% of women; chi-squared test, $p < 0.001$). We then calculated the percentage of women in each of these majors as well as the frequency of occurrence of gendered mismatches between prior GPA and student grade. We also examined the number of STEM units across gender and in relation with course grade outcomes.

RQ3: How are Other Demographic Factors Associated with Student Course Outcomes and Gendered Equity Gaps?

Students hold multiple identities, many of which are not easily visible nor reportable (Chaudoir and Quinn, 2010), that may influence their college experiences, and thus, course outcomes. To assess what demographic variables other than gender may relate to performance (i.e., final course grade) and interact with gender in this upper-division physiology course, we created a new linear multilevel model using the "lme4" package in R (Bates *et al.*, 2015) that included the following variables as fixed effects: prior GPA, gender, PEER status, low socioeconomic status, first-generation college student status, English as Second Language (ESL) classification, course quarter (Fall, Winter, or Spring) and admission level (whether they began college as a freshman or a transfer student). In addition, our model included a random effect to control for the clustered nature of students in our data within course offerings (Theobald, 2018). We did not include a random effect for individual students because there were no repeated measures in our dataset (we only used grades from the first term taken for all students). In all models, the dependent variable was final course grade, transformed from a letter grade to a numeric 4.0 scale, following the university registrar's policy for GPA calculation (A+ = 4.0, A = 4.0, A– = 3.7, B = 3. B– = 2.7, C+ = 2.3, C = 2, C– = 1.7, D+ = 1.3, D = 1, D– = 0.7, F = 0). The global multilevel model was as follows:

$$CourseGrade \sim \beta_{PriorGPA} + \beta_{Gender} + \beta_{PEER} + \beta_{FirstGen} + \beta_{ESL} + \beta_{LowSES} + \beta_{AdmitLevel} + \beta_{Quarter} + (1 \mid Offering) + \varepsilon$$



Systemic Advantage Index (SAI)	0	1				2						3				4
Man		X				X			X		X	X	X	X		X
Non-PEER			X			X	X			X		X	X		X	X
Continuing Generation				X			X	X			X	X		X	X	X
Mid to high SES					X			X	X	X			X	X	X	X

FIGURE 1. Systemic advantage index (SAI) values for students with access to a range of advantages in the higher education system conferred by student gender, race/ethnicity, first-generation college-going status, and socioeconomic status. Gray cells marked with X represent an advantage available to that student. The top row indicates the SAI index value for each theoretical student.

Starting with this global model, we used the package "MuMIn" (Barton, 2020), to assess the fit of all possible models that included various subsets of the fixed effects (no interactions). We compared model fit using Akaike's Information Criterion (AIC) (Akaike, 1974), Bayesian Information Criterion (BIC) (Schwarz, 1978), model log-likelihood and weights. We used AIC rather than AICc (i.e., AIC corrected for small sample size; Burnham and Anderson, 2004) as our model selection criterion because our ratio of sample size to model parameter number (max = 12) was sufficiently high (a ratio > 40 is recommended for use of AIC instead of AICc) (Burnham and Anderson, 2004). In addition, we included BIC values in our analysis, as it is more conservative and accrues larger penalties for adding variables to the models (Dziak *et al.*, 2012).

We initially found five models that had AIC values within two points from the best-fit model (Supplemental Table S5), indicating they may fit the data similarly to the lowest weighted model (Burnham and Anderson, 2004). The model averaged coefficients, where each coefficient is weighted by the Akaike weight of the model, for the top five models, can be found in Supplemental Table S6. As the models with the lowest AIC and BIC were not the same, we selected the model using the lowest BIC and log-likelihood as our criteria (bold row, Supplemental Table S5), as these criteria are more conservative to adding additional variables than AIC and thus allow us to select the simplest model with the highest degrees of freedom (Dziak *et al.*, 2012). Additionally, the lowest AIC model included extra variables that did not have significant coefficients when all viable models were averaged. This best fit model with main effects only included prior GPA, gender, PEER status, first-generation status, and low socioeconomic status as main effects (Model I, Results section).

As students' educational experiences are associated with multiple axes of their identities (Crenshaw, 1991; Pearson *et al.*, 2022), we hypothesized that student outcomes in the course could also be associated with specific combinations of identities such as being a woman and PEER. Given the mismatch we found between student prior GPA and gender, we also expected that different axes of student identity might moderate the association between prior GPA and course outcomes. To assess this possibility, we evaluated whether adding each possible two-way interaction between the main effects in our previous best model (Model I) improved model fit by comparing AIC, BIC, log-likelihood values, and weights across 10 models. AIC/BIC tables (Table 4) were produced using the "bbmle" package (Bolker, 2022). Upon comparing the AIC, BIC and log-likelihood of these 10 models, only one model had the lowest AIC, BIC, and log-likelihood score (Table 4). We thus moved forward with this model as the best fit model including second-order interactions, which contained prior GPA, gender, PEER status, first-generation status, low socioeconomic status and an interaction between prior GPA and gender (Model II, in Results).

For both selected models (Model I and Model II), we report the robust coefficients in the paper. A comparison of the robust and nonrobust coefficients and the final robust weights are shown in Supplemental Tables S7 and S8.

Systemic Advantage Index

Within the current landscape of STEM higher education, specific demographic characteristics, such as being white (Eagan *et al.*, 2010; Chang *et al.*, 2014; McGee, 2020) or a man (Cimpian *et al.*, 2020; Perez-Felkner, 2018) have been repeatedly shown to be associated with better academic outcomes (Suárez *et al.*, 2021).This trend is also observed for continuing generation students (Bettencourt *et al.*, 2020) as well as having a middle to high socioeconomic status (Niu, 2017). These advantages likely have a cumulative effect that is independent of the specific demographic characteristic that confers the advantage. Using this framework, one can divide the student population into five groups (Figure 1) depending on their total number of systemic advantages (Castle, 2021). For example, students who are women, first-generation, from a low socioeconomic status and PEER have zero systemic advantages and thus would be in the systemic advantage index 0 group (SAI = 0), while a white, continuing generation, high socioeconomic status man would be in the SAI = 4 group, with four systemic advantages. Students who are assigned one to two advantages may have any combination of advantages, meaning that a white, first generation, low socioeconomic status, man would have an SAI = 2, but so would a PEER, woman who was from a middle/high socio-economic status and continuous generation (also SAI = 2). This framework does not imply that all students with any two advantages are equivalent; the lived experiences of the two recent example students with SAI = 2 will undoubtedly differ, and the specific experiences of each individual student will impact their educational trajectory differently. Still, the SAI framework proposes that in the current higher education landscape, a student with three advantages is likely to have better academic outcomes than a student with only one or two advantages, regardless of the specific advantage. We did not include gender in our final SAI assignment as we wanted to examine interactions between student gender and other demographic factors (so SAIs used will range from 0 to 3). We acknowledge that while this approach allows us to examine potential interactions between student identities and gender, the advantages conferred by these identities do, of course, not necessarily lead to the same outcomes or experiences for all students.

To explore the relationship between systemic advantages and course outcomes, we plotted the average course grade for men and women across all course offerings against SAI. We then used two-tailed *t* tests to compare average grades and average prior GPAs for men and women at different levels of SAI.

TABLE 2. Items from the introductory questionnaire included in the analysis

Statement	Category
I am a scientist.	Science identity[a]
I feel like I belong in the field of science.	
In general, being a scientist is an important part of my self-image.	
I often find myself feeling more anxious than my classmates about how I am doing in my class.	Course anxiety
I usually do better in my science courses than in my GE (general education) courses.	Science self-efficacy (SSE)

[a]Science identity items are from a validated scale used in (Robnett *et al.*, 2015).

RQ4: What Environmental and Affective Factors may be Associated with the Observed Gendered Performance Gaps?

Previous studies have shown that the gender composition of the student body and the instructional team can affect student performance (Carrell *et al.*, 2010; Eddy *et al.*, 2014; Bailey *et al.*, 2020). We, therefore, evaluated whether the gender composition of the class and instructional team related to final grades and interacted with gender in upper-division human physiology. To do this, we extended the best fit model (Model II; Results) to include two additional characteristics as fixed effects: percent of women in each individual class offering (*PercentWomen*) and presence of a woman on the instructional team in each offering (as a dichotomous categorical variable: *WomanProfessorPresent*). Given prior research (Bailey *et al.*, 2020), we also evaluated whether including each of these fixed effects with an interaction with student gender (e.g., *PercentWomen*Gender, WomanProfessorPresent*Gender*) improved model fit. We compared model fits using AIC, BIC, and log-likelihood information criteria.

Because many previous studies have shown overall class size can negatively relate to women's course outcomes (Ballen *et al.*, 2019; Bailey *et al.*, 2020; reviewed in Odom *et al.*, 2021), we also evaluated whether class size was correlated with the gender disparities we observed in course grade.

Introductory Course Questionnaires to Examine Affective Factors

Gendered differences in affective factors, such as self-efficacy, science identity, and test anxiety have been observed in introductory courses (Eddy and Brownell, 2016; Ballen *et al.*, 2017), and have been found to relate to grade outcomes (Ballen *et al.*, 2017). During the Fall 2018 through Spring 2019, students taking the upper physiology course completed a large questionnaire during the first week of class that included items pertaining to a variety of psychosocial factors, including science identity, self-efficacy, mental health, perceptions of discrimination and micro-affirmations, and more. For this study, we selected items *a priori* that might pertain to gendered differences in science identity, science self-efficacy, and course anxiety from the larger questionnaire (reviewed in Eddy and Brownell, 2016). Table 2 lists the questionnaire items we used and indicates, where applicable, the relevant source in the research literature. The student data in these 3 quarters showed similar best-fit model estimates (Model II, Results section) for prior GPA, gender, and the interaction to those of the broader, 10-year dataset (see Methods for RQ3 above for modeling details; model coefficients for the data from these 3 quarters is shown in Supplemental Table S9).

To analyze our questionnaire data, we first filtered the initial questionnaire responses by consent to be included in the study and by time taken to complete the questionnaire (students who took <3 min were removed). To validate student attention, we added the following item among those relevant to this study: "*Just to make sure you are reading the questions, please select 'Neutral'.*" Students who did not answer "Neutral" to this question were removed from the dataset. For students who completed the questionnaire more than once, we only used their first attempt. After filtering the responses, questionnaires from 896 students remained, representing 81, 97 and 88% of students with course grade data across the three quarters, respectively. Likert scale questions were numerically coded on a scale from –2 ("Strongly Disagree" or "Never") to 2 ("Strongly Disagree" or "Always") with 0 representing "Neutral." The science-self efficacy item *("I usually do better in my science courses than in my GE [general education] courses"*) and the course anxiety item ("*I often find myself feeling more anxious than my classmates about how I am doing in my class*") were both created for this introductory questionnaire. To create an average science identity score, we averaged the responses across three survey items: "*I am a scientist.*", "*I feel like I belong in this field of science.*", and "*In general, being a scientist is an important part of my self-image*". These items are a subset of a previously-validated, five-item science identity scale (Robnett *et al.*, 2015). Confirmatory factor analysis, run using the "lavaan" package in R (Rosseel, 2012), showed that these three science identity items loaded onto a single factor matching the original factor structure (Robnett *et al.*, 2015). Model fit for this factor also met recommended fit criteria (comparative fit index >0.95, standardized root-mean-square- residual [SRMR] <0.08) (Hu and Bentler, 1999).

For the average science identity score, we examined average differences between men and women using two-sample *t* tests. For the single-item factors, such as course anxiety ("*I often find myself feeling more anxious than my classmates about how I am doing in my class.*") and science self-efficacy ("*I usually do better in my science courses than my GE courses*"), we examined differences between genders using Wilcoxon rank-sum tests.

As course anxiety, science identity and self-efficacy have been shown to be associated with students' overall academic performance (Williams and George-Jackson, 2014; Ballen *et al.*, 2017; Lent *et al.*, 2018), we also evaluated whether students' responses in these initial questionnaires correlated with their final course grade. As we found significant correlations between grades and questionnaire item responses for both men and women, we built linear models with final course grade predicted by each affective factor item score, gender, and prior GPA. We used model comparison to compare models including the main effects only (*CourseGrade ~ Prior GPA + ScienceIdentity + ScienceSelfEfficacy + CourseAnxiety + Gender + (1|Offering)*), main effects and an interaction between prior GPA and gender (as in Model II), and a model building on the latter that included interactions between gender and all affective factors. No other demographic variables were added to the model because our research questions revolved around the influence of affective factors on gender gaps specifically and because the sample size of the questionnaire data was more limited. This approach allowed us to examine whether controlling for prior academic performance removes the relationship between any introductory questionnaire item responses and course grade, or if a relationship persists even when we include prior GPA and its interaction with gender. We report the model estimates from the linear model with the lowest information criteria (AICc, BIC, and log-likelihood). We used AICc instead of AIC as an information criterion because the sample for the questionnaire data was much smaller than that for the full dataset. In addition to this linear model approach, we also tested the partial mediation of affective factor survey scores on the relationship between prior performance (prior GPA) and course grades. Partial mediation analyses can be found in Supplemental Figure S2.

RESULTS

RQ1: To What Extent are Gendered Equity Gaps, Like Those Seen in Introductory Biology Courses, Present in Large Upper-Division Biology Courses?

Across 35 course offerings encompassing 13,184 students, men had significantly higher average course grades than women in the upper-division human physiology course under study (Figure 2A; 2.78 ± 0.014 vs. 2.55 ± 0.010, respectively; two-sample t test, $p < 0.001$). Further, this gendered performance gap was larger than that observed in the university's introductory biology course (Figure 2B, difference in average grades in lower division (W–M): –0.133 ± 0.022, upper division –0.221 ± 0.024) for the subset of students who took the course in the same institution ($n = 9{,}735$; 74% of total sample; a gap which itself was comparable with gender gaps reported for lower-division biology courses at other institutions; Eddy and Brownell, 2016; Odom *et al.*, 2021). Examination of the distribution of final letter grades (Figure 2C) showed a significant disparity in the percentage of men and women who received As in the course (24% vs. 17%, respectively; *chi-squared* = 70.3, $p < 0.001$). In addition, 10% of women and 7% of men received failing grades (Ds or Fs) (*chi-squared* = 38.6, $p < 0.001$). These gendered performance gaps were consistently observed across all 10 years of data examined (2009–2019; Figure 2D)

Linear multilevel models consisting of course grade as outputs and gender as the only independent variable and controlling for variation between offerings (see Methods), show that gender explains a significant amount of the variance in course grade, with being a woman having a negative coefficient (Figure 2E; $\beta = -0.23 \pm 0.02$; $p < 0.001$). This model predicts that in this course, women, on average, earn a class grade of about 0.23 points lower (on a 4.0 grade scale) than men.

RQ2: How do Academic Factors, such as Prior GPA, Major, and Number of STEM Units Taken, Relate to Student Outcomes and Gendered Equity Gaps?

Across the population of students, we found no significant difference in prior college GPA between men and women (3.07 ± 0.005 vs. 3.09 ± 0.007, respectively, two-tailed t test; $t = 1.69$, $p = 0.091$). Adding prior GPA as a control for prior academic performance to our initial linear multilevel model had a minor impact on the association of gender and course outcomes ($\beta = -0.19 \pm 0.01$; $p < 0.001$; Figure 2E). To examine if the association of prior GPA with course outcome varied with gender, we added an interaction effect between gender and prior GPA. In this simple model, the interaction was significant ($\beta = 0.12 \pm 0.03$; $p < 0.001$) and it resulted in a nearly three-fold increase in the impact of being a woman on predicted grade ($\beta = -0.56 \pm 0.08$; $p < 0.001$) (Figure 2E, right bar). This result means that the association of gender with course outcomes was moderated by prior GPA, with the impact of gender being stronger for women with lower prior GPAs. For instance, this model predicts that among students with a prior GPA of 2.0, women would receive a grade approximately 0.32 points lower than men (out of 4.0), while for those with prior GPA of 4.0, the difference would be approximately 0.09 points (for further examples, see Supplemental Table S10 and Supplemental Figure S1). Thus, our model predicts that when students enter this course with identical prior GPAs, a woman will earn a lower grade than a man, regardless of her prior GPA, the latter modulating *how much lower* her grade will be.

To better visualize the relationship between disparities in prior GPA and disparities in course outcomes, we plotted the average grade disparity by gender against the average prior GPA disparity by gender for each individual course offering (Figure 2F). We found that women earned lower grades (on average) than men in all but one offering of the class across the 10 years examined. Furthermore, in 46% ($n = 16$) of these offerings, women entered the course with an average prior GPA that was *higher* than that of men in the class, yet still earned lower grades, on average (Figure 2F; gray quadrant).

We then examined whether these gendered preparation-outcome mismatches varied across the top six most-populous majors of the course, as these majors varied in their STEM requisites (which could impact GPA and academic preparation, see Methods) as well as their percentage of women. We found mismatches between average prior GPA and average course grades across the top 6 majors (range 6–29% of offerings), despite differences in major STEM requirements and percentage of women enrolled in the course (range: 51–85% across these majors). For instance, the Neurobiology, Physiology and Behavior major and the Psychology major had the same number of mismatched offerings (Figure 3; B and D; 10 offerings, 29%), despite having a 9% difference in percentage of women in the course (64 and 73% respectively). Similarly, the Animal Science major and the Biochemistry and Molecular Biology major both had 20% of offerings where the mismatch occurred (Figure 3; C and E), while the percent women differed in 33% (84 and 51%, respectively). Thus, gendered mismatches between prior academic performance and course grades are present in majors from various disciplines. In addition, the percentage of mismatches did not significantly correlate with the number of women in the major that were taking the course (Pearson's $r = -0.34$, $p = 0.499$).

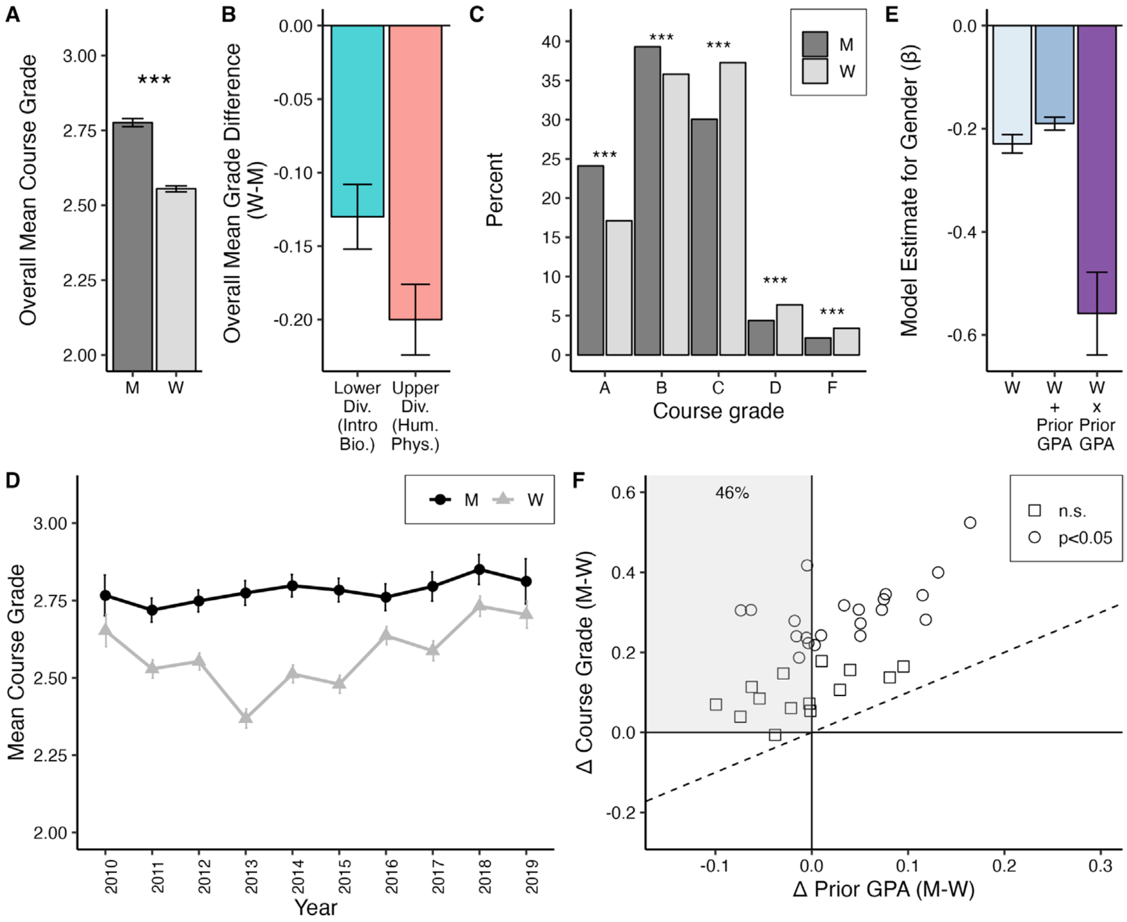


FIGURE 2. Performance gaps associated with gender in upper-division physiology are greater than those in lower division introductory biology and what might be expected based on prior academic performance. (A) On average, men (dark) received significantly higher course grades compared with women (light) across all offerings of human physiology (two-sample t test, $p < 0.001$, $n = 13{,}184$). (B) The overall average difference in course grades between men and women was larger in upper-division physiology compared with lower division introductory biology for students for whom grade data were available for both courses (difference in average grades in lower division (W–M): −0.13 ± 0.02, upper division −0.22 ± 0.02, $n = 9{,}735$). (C) When examining actual letter grades received, men were more likely to receive As and Bs than women, but less likely to receive Cs, Ds, and Fs (chi-squared test of independence, all $p < 0.001$). (D) Across the 10-year study period, men (dark circles) consistently showed higher grades on average than women (light triangles). Each point is the average of 3–5 offerings of the course. (E) The model-estimated effect of being a woman on course grade is largest when an interaction between prior overall GPA and being a woman is included. Bars represent model estimates and standard errors produced by multilevel models including only the fixed variables listed on the x axis, along with course offering as a random effect. (F) In all but one offering (each circle/square is an offering), women received lower average grades than men (data in top half of plot). In 46% of course offerings, women received lower grades on average than men, but entered the course with a higher prior GPA on average (upper-left gray quadrant); of these, eight offerings had a significant difference in grades between men and women (two sample *t* test, $p < 0.05$). There were no offerings where men entered the course with a higher average prior GPA and left the course with a lower grade than women (bottom-right quadrant). The dashed line represents the 1–1 line. Circles represent offerings where men and women's average grades were significantly different ($p < 0.05$), while squares represent offerings where the observed difference was not statistically significant. In all cases, error bars represent standard error of the mean (SEM). In figure: *** $p < 0.001$.

Because students can still take STEM courses even if they are not required for their majors, and many prehealth students take additional STEM courses to satisfy graduate school admissions requirements, we also examined the number of STEM units taken before the course for each student in our study. We did not find a significant difference in the average number of prior STEM units between men and women (63.9 vs. 64.3 units respectively; two-sample *t* test, $p = 0.476$). Further, course grade was not significantly correlated with prior STEM units (Pearson's $r = –0.01$, $p = 0.373$).

RQ3: How are Other Demographic Factors Associated with Student Course Outcomes and Gendered Equity Gaps?

Using demographic data for our student dataset from the university registrar (Table 1), we applied multilevel modeling and model selection to examine the possible association of other demographic factors, in addition to gender, with outcomes in upper-division physiology. We also included a random effect in the model to account for any variability between course offerings (see Methods). Model selection upon a global model that

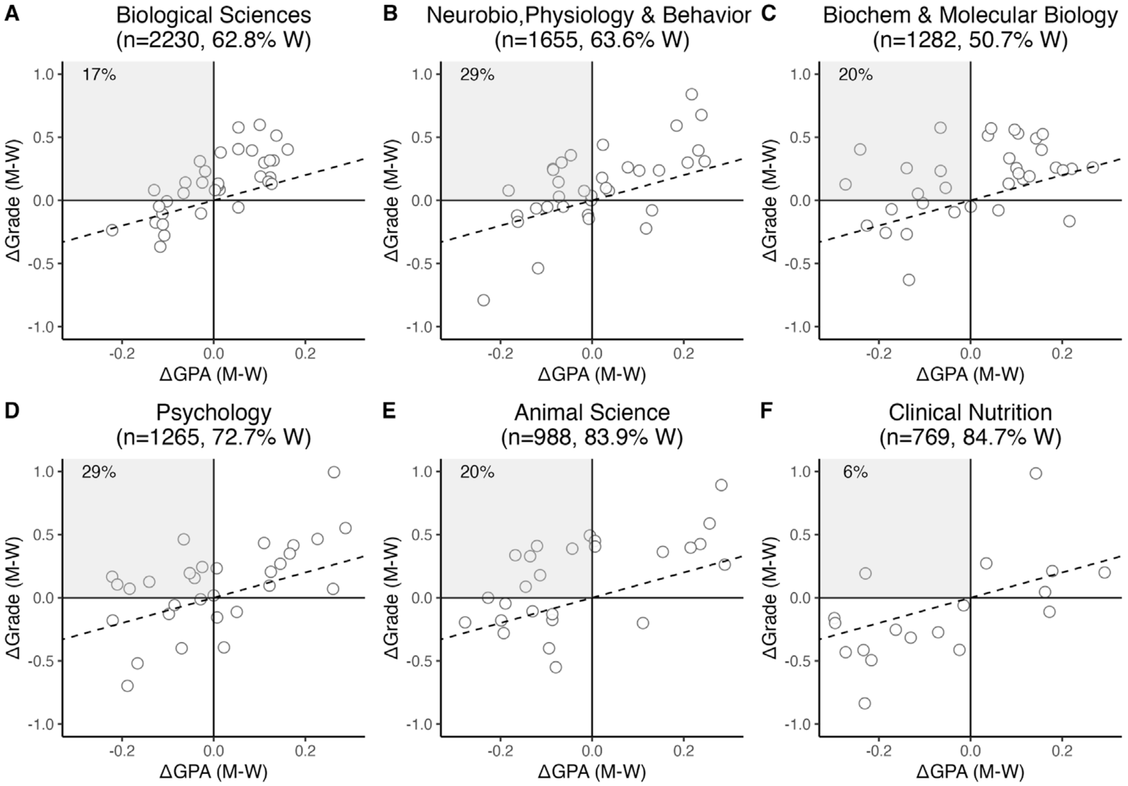


FIGURE 3. Gendered differences in prior GPA and course grade vary across the top six most populous majors in the course. Plots show average course grades and average prior GPA for each offering, disaggregated by the top six majors, ranked by number of students who had declared that majors at the time of course enrollment. In each plot, the top half of the plot represents offerings where men had a higher average course grade than women, while the bottom half represents offerings where women received higher average course grades than men. The right half represents offerings where men entered the course with a higher prior GPA than women on average, while the left half represents offerings where women entered the course with higher prior GPA than men. The dashed line represents the 1–1 line. "Preparation-outcome Mismatches" occur when offerings fall in the bottom right quadrant (men enter the course with higher prior GPA on average but receive lower average grades) and in the top left quadrant (women enter the course with a higher prior GPA but receive lower average grades than men on average). The percent of offerings that fall into this latter "mismatch" is annotated in the upper left. Majors shown in panels A–C are housed in the Biological Sciences, while the other majors (D–F) are housed in other colleges. Circles represent offerings of the course. Sample sizes for each major (*n*) and the percentage of students in that major that are women (%W) are reported above each graph.

included gender, prior GPA, PEER status, first-generation status, socioeconomic status, English as second language (ESL) status, course quarter, transfer student status as main effects and course offering as a random effect yielded the following best-fit model:

Model I. *Main effects best fit model:*

$$CourseGrade \sim \beta_{PriorGPA} + \beta_{Gender} + \beta_{PEER} + \beta_{FirstGen} + \beta_{LowSES} + (1 \mid CourseOffering) \cdot + \varepsilon$$

The robust estimates for all demographic variables retained in the best model (Model I) were negative, indicating that being a woman, PEER, first-generation student, or from a low socio-economic status (*LowSES*), was associated with earning lower grades in this upper-division course (Table 3). All main effect estimates were statistically significant except for socio-economic status, yet the model selection process included this variable in the best model as it improved the fit of the final model. Specifically, our model predicted that when comparing students with identical prior GPA a student who was a woman, PEER, first-generation and low socioeconomic status would, on average, earn a grade that was about 0.37 points lower (on a 4.0 grade scale) than a man who was not-PEER, continuing generation, and was not from a low

TABLE 3. Robust[a] model estimates for Model I, the best fit model including only main effects

Variable	Estimate(β)	Std.Error	*p*[b]
(Intercept)	–0.932	0.043	**<0.001**
PriorGPA	1.123	0.012	**<0.001**
Gender	–0.189	0.012	**<0.001**
PEER	–0.049	0.015	**<0.001**
FirstGen	–0.070	0.013	**<0.001**
LowSES	–0.021	0.014	0.145

[a] Robust estimation was used since the residuals from Model I did not meet assumptions of normality and homoscedasticity, and this method is more robust to outliers in the data. (See *Methods* for more details)

[b] *p*-values that were significant ($p < 0.05$) are shown in bold.

socioeconomic status. For example, if both students entered with a 3.0 GPA, the model predicts the former student would earn a grade of 2.35, while the latter would earn a grade of 2.76, leading to a difference of a full plus/minus grade (C+ vs. a B–, respectively).

Adding a variable that represents prior academic performance when modeling student course outcomes provides valuable information about the factors that are associated with how students perform in a course. However, this practice can feed a student deficit narrative (Smit, 2012), where student underperformance is considered to just be the result of lack of academic preparation, instead of a confluence of institutional, structural, pedagogical, social, and academic factors. To examine how including prior GPA affected the relationship between student identities and course outcomes, we compared coefficient estimates for demographic variables between our best fit main effect model and an identical model lacking prior GPA (Supplemental Table S11). Exclusion of prior GPA increased the negative coefficients for socio-economic, low impact and first-generation status, while having little effect on the gender coefficient (Supplemental Table S11).

To assess how multiple axes of student identities interact with each other, as well as with prior GPA to influence course outcomes, we created a series of new models, each of which consisted of the main effects from Model I and an interaction term between two of these main effects (Table 4). Of the 10 models created, the best fit model included prior GPA, gender, first-generation status, PEER status and low socioeconomic status as predictors that best explained the variation in course grades, as well as an interaction between prior GPA and gender (Model II). None of the other two-way interactions improved the model fit (Table 4; all had AIC values within 2 of the best-fit model without interactions [Model I], suggesting that these models were not meaningfully different from Model I [Burnham and Anderson, 2004]). Table 5 shows the robust regression coefficients for Model II.

Model II. *Best fit model including interactions:*

$$CourseGrade \sim \beta_{PriorGPA} + \beta_{Gender} + \beta_{PEER} + \beta_{FirstGen} + \beta_{LowSES} + \beta_{PriorGPA*Gender} + (1|CourseOffering) + \varepsilon$$

The addition of the interaction between prior GPA and gender resulted in a large change to the coefficient for gender (shifting from –0.19 to –0.54) compared with the model with main effects only. However, with an interaction now present, the main effects of prior GPA and gender cannot be interpreted individually as having direct effects on course grade as the variables are now conditional on each other. For instance, the coefficient for gender would indicate a reduction of –0.54 in course grade for a woman whose prior GPA was zero, an impossible circumstance in an upper-division course. Our updated model predicts that a student who was a woman, PEER, first-generation and low socioeconomic status (student W_{PFGLS}) would earn a grade that was about 0.24 to about 0.68 lower (on a 4.0-point scale) than a man who was neither PEER, nor first generation nor low socioeconomic status (student M), depending on their prior GPA. As an example, if both these students entered with the same incoming GPA (example GPA=2.1), the model predicts that the student M would receive a C– grade (1.72), allowing them to pass the course, while student W_{PFGLS} would receive a D+ (1.28), a non passing grade. More examples are shown in Supplemental Table S12.

Systemic Advantages

As students belong to multiple nonmutually exclusive demographic identities, any of which may affect their experience in a course and course outcomes by itself or through interactions with others, we modeled multiple demographic variables using a systemic advantage index (SAI) as described by Castle (2021) and Whitcomb *et al.* (2021). In this approach, we assigned each student an index that sums the number of advantages conferred to that student in higher education by belonging to a historically privileged racial/ethnic group (i.e., non-PEER race/ethnicity),

TABLE 4. Comparison of models including pairwise interactions across all main effects

Model	df	AIC	dAIC[a]	BIC	dBIC[a]	LogLik	dLogLik[a]
Top model (MODEL I) **(main effects only):** CourseGrade ~ PriorGPA + Gender + PEER + FirstGen + LowSES + (1∣Section)	8	28662.5	0.0	28722.4	0.0	–14323.3	0.0
+ **PriorGPA*Gender**[b]	**9**	**28644.6**	**–17.9**	**28712.0**	**–10.4**	**–14313.3**	**9.9**
+ PriorGPA*FirstGen	9	28661.6	–0.9	28729.0	6.6	–14321.8	1.4
+ PriorGPA*LowSES	9	28662.7	0.1	28730.0	7.6	–14322.3	0.9
+ PriorGPA*PEER	9	28664.1	1.6	28731.5	9.1	–14323.0	0.2
+ Gender*PEER	9	28663.8	1.3	28731.2	8.8	–14322.9	0.3
+ Gender*FirstGen	9	28661.8	–0.7	28729.2	6.8	–14321.9	1.4
+ Gender*LowSES	9	28664.5	2.0	28731.8	9.4	–14323.2	0.0
+ PEER*FirstGen	9	28662.8	0.3	28730.1	7.7	–14322.4	0.9
+ PEER*LowSES	9	28664.2	1.7	28731.6	9.2	–14323.1	0.1
+ FirstGen*LowSES	9	28664.5	2.0	28731.8	9.4	–14323.2	0.0

[a] Information criteria (AIC, BIC, log–likelihood) for all models are compared with the base model with no interactions, selected by comparing models using AIC and BIC from a global model which included all possible fixed effects.

[b] The model in bold includes an interaction between PriorGPA and Gender that significantly improved the base model with main effects only (i.e., lowered the AIC and BIC values) (Burnham and Anderson, 2004).

TABLE 5. Robust model estimates for Model II, the best fit model including interactions

Variable	Estimate(β)	Std.Error	p^a
(Intercept)	–0.708	0.066	**<0.001**
PriorGPA	1.158	0.020	**<0.001**
Gender	–0.540	0.079	**<0.001**
PEER	–0.048	0.015	**0.002**
FirstGen	–0.068	0.013	**<0.001**
LowSES	–0.021	0.014	0.130
PriorGPA*Gender	0.113	0.025	**<0.001**

[a] *p*-values that were significant ($p < 0.05$) are shown in bold.

being from moderate to high socioeconomic status, and being a continuing generation college-going student (see Figure 1 and Methods for more details). This systemic advantage framework provides a way to explore intersections between student identities and to examine systemic barriers in higher education (Castle, 2021).

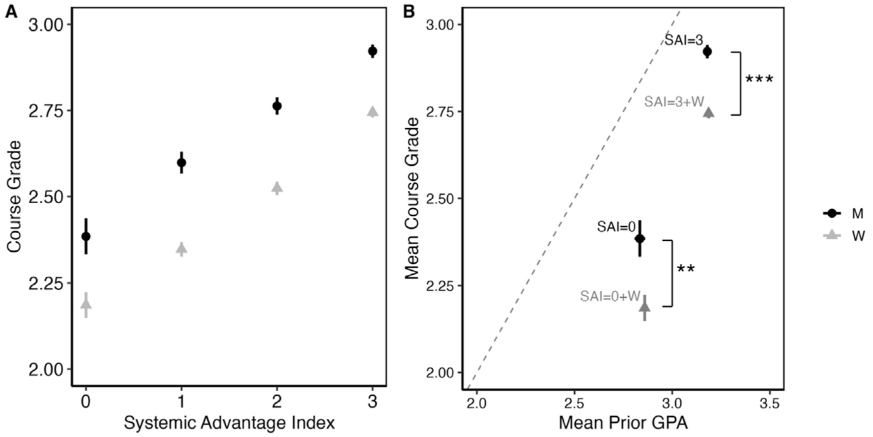


FIGURE 4. Gendered performance gaps are present across a range of student systemic advantages based on demographic identities. Different identities, related to the main effects in the best-fit model (first generation status, PEER status, and low socioeconomic status) were considered "advantages" and used to generate a "systemic advantage index" (SAI) as described in Castle, 2021; Whitcomb *et al.*, 2021. If a student has access to all the advantages conferred by being continuing-generation college-going (i.e., not first-generation), non-PEER, and not from a low socioeconomic status, the student is considered to have an SAI = 3. On the other hand, students who are first-generation, PEER, and from low socioeconomic backgrounds would be considered to have none of these advantages, so would have an SAI = 0. Students with access to any one or two of these advantages would have SAI = 1 or SAI = 2, respectively (see Figure 1 for details). (A) Average grade is shown for both men (black) and women (gray) across the spectrum of SAI identities. Men's grades were significantly higher than women's for all SAI values (two-tailed *t* tests, $p < 0.001$ for SAI = 1–3; $p < 0.01$ for SAI = 0). Error bars represent standard errors of the mean (SEM). (B) Significant gendered differences in mean grade, but not mean prior GPA, are present at both ends of the SAI spectrum. Here, "SAI = 0 + W" indicates a woman with none of the listed advantages, where SAI = 0 is a man with none of the listed advantages. On the other end of the spectrum, "SAI = 3" is a man with access to all listed advantages whereas "SAI = 3+W" is a woman with all those advantages. Women received significantly lower average grades than men when they had no listed advantages (2.19 ± 0.04 vs. 2.38 ± 0.05, respectively; two-sample *t* test, $p < 0.01$) and when they had access to all advantages (2.74 ± 0.02 vs. 2.92 ± 0.02, respectively; two-sample *t* test, $p < 0.001$). Prior GPA did not significantly differ with gender at either end of the spectrum (no other advantages, men: 2.83 ± 0.01, women: 2.86 ± 0.02, two-sample *t* test, $p = 0.435$; all other advantages, men: 3.18 ± 0.03, women: 3.19 ± 0.01, two-sample *t* test, $p = 0.611$). Error bars represent standard error of the mean (SEM). In figure: ** $p < 0.01$; *** $p < 0.001$.

Across the spectrum of systemic advantages, men received higher grades than women regardless of how many other systemic advantages they had (Figure 4A; two-tailed *t* tests, $p < 0.001$ for SAI = 1–3; $p < 0.01$ for SAI = 0). Again, this gendered difference appears driven by course grade rather than differences in prior GPA. Men received significantly higher grades than women when neither of them had any other systemic advantages (Figure 4B; bottom; two-tailed *t* test $p < 0.01$) and when they both had all other systemic advantages (Figure 4B; top; two-tailed *t* tests, $p < 0.001$), even though in both cases, the prior GPA did not significantly differ between men and women in each group (two-tailed *t* tests, all $p > 0.05$).

RQ4: What Environmental and Affective Factors may be Associated with the Observed Gendered Performance Gaps?

Instructor gender, gender representation in the classroom, and class size. To examine if instructor gender and percentage of women in a class interacted with student gender to influence course and exam scores in our course, we modified our best model including interactions (Model II) to add the presence of a woman on the instructional team and/or the percentage of women in each class offering as main effects, as well as an interaction of each of these variables with student gender (Table 6). When we compared model fit, neither of these main effects nor any of the interactions between these variables and student gender improved fit over that of the previous best fit model (Model II). Further, neither the estimates of these variables, nor their interaction with student gender, were significant (Table 6). We also examined the effect of class size on the gendered grade disparity by calculating the correlation between class size (range: 177–528 students) and the differences between men's and women's grades across all offerings and found no significant relationship (Pearson's correlation, $r = –0.038$, $p = 0.828$).

Affective factors. We used data from questionnaires administered at the beginning of three course offerings ($n = 896$ students; see Methods) to evaluate if gendered differences in affective factors such as science identity, science self-efficacy, and course anxiety were present in upper-division students before they engaged significantly with course material. Correlations between student-reported scores on these questionnaire items and course grades were also calculated to assess if these affective factors were associated with student outcomes (Figure 5).

Women had significantly lower levels of science identity than men upon entry to

TABLE 6. Model selection and robust model estimates for models including environmental classroom variables

Model	df	AIC	dAIC[a]	BIC	dBIC[a]	logLik	dLogLik[a]	Robust model estimates (β ± SE)			
								Gender[b]	Percent Women	Woman Professor Present	Interaction with Gender
[Best fit]	9	28644.6	0.0	28712.0	0.0	–14313.3	0.0	–0.57 ± 0.09			
+ Percent Women	10	28646.1	1.5	28721.0	8.9	–14313.1	0.3	–0.54 ± 0.08	–0.002 ± 0.005		
+ WP[c] Present	10	28646.6	1.9	28721.4	9.4	–14313.3	0.0	–0.54 ± 0.08		0.025 ± 0.027	
+ Percent Women + WP[c] Present	11	28647.9	3.2	28730.2	18.2	–14312.9	0.4	–0.54 ± 0.08	–0.001 ± 0.003	0.027 ± 0.029	
+ Percent Women* Gender	11	28644.8	0.1	28727.1	15.1	–14311.4	1.9	–0.83 ± 0.21	–0.003 ± 0.003		–0.004 ± 0.003[#]
+ WP[c] Present * Gender	11	28648.4	3.8	28730.8	18.7	–14313.2	0.1	–0.53 ± 0.08		0.034 ± 0.032	–0.013 ± 0.025

[a] Information criteria (AIC, BIC, log–likelihood) for all models are compared with the best fit model (Model II) which includes Gender, PEER, FirstGen, LowSES, and an interaction between PriorGPA*Gender.
[b] All robust estimates for the effect of Gender were significant at the $p < 0.001$ level in all models.
[c] WP: Women Professor.
[#] $p = 0.069$

the course (two-tailed t test; $t = –3.26$, $p = 0.001$; Figure 5A). Further, the median response to the science self-efficacy item (SSE) for women was lower than that for men, indicating more disagreement with the statement (Wilcoxon's rank-sum: $p < 0.001$; Figure 5B). In response to the question "*I often find myself feeling more anxious than my classmates about how I am doing in my class*", the median response for women was higher than for men indicating that women entered the course with more anxiety than men (Figure 5C; Wilcoxon's rank-sum: $p < 0.001$).

All questionnaire items significantly correlated with final course grades for women (Figure 5D, Pearson's correlations), with course anxiety being negatively correlated with grade ($r = –0.21$, $p < 0.001$) and SSE and science identity being positively correlated with grade ($r = 0.32$, $p < 0.001$ and $r = 0.18$, $p < 0.001$, respectively). For men, only SSE and science identity were significantly correlated with grade, with the correlation being positive ($r = 0.26$, $p < 0.001$ and $r = 0.12$, $p = 0.001$, respectively; Figure 5E). Course anxiety was not significantly related to grade for men ($r = –0.07$, $p = 0.289$).

To examine how each of these affective factors was associated with student grades while controlling for prior GPA, we conducted model selection on a linear model that initially included prior GPA, each affective variable, gender, and all possible interactions between gender and the affective variables (Supplemental Tables S13, and S14). We found that the model that best fit the questionnaire data included main effects for all affective factors, gender, and prior GPA, as well as interactions between gender and prior GPA and gender and course anxiety (Table 7). All coefficients included in the model were statistically significant ($p < 0.05$) except for the interaction between gender and anxiety, which despite improving model fit, only showed a trend towards significance ($p = 0.077$).

This model was:

Model III. *Best fit model for affective factors (questionnaire data).*

$$\begin{aligned} CourseGrade \sim\ & \beta_{PriorGPA} + \beta_{Gender} + \beta_{ScienceIdentity} \\ & + \beta_{ScienceSelfEfficacy} + \beta_{CourseAnxiety} \\ & + \beta_{Prior\ GPA*Gender} + \beta_{Gender*CourseAnxiety} + \varepsilon \end{aligned}$$

Consistent with the correlation findings, science identity and SSE scores were positively associated with course grades, even when controlling for prior GPA. The exclusion of the interaction effect for these affective factors and gender from the best model (as they did not improve model fit, see Supplemental Tables S13 and S14), indicates that the association between science identity and SEE and course grades was comparable for both genders. Because the best model included an interaction between gender and course anxiety, it is not possible to interpret the coefficient value or significance of the fixed effect for anxiety in isolation. The negative coefficient for the interaction between course anxiety and gender indicates that for women, increasing scores on the course anxiety questionnaire item were negatively associated with course grade, even when controlling for prior GPA and other affective factors. This model predicts that if two women have the same prior GPA (e.g., 3.0), science identity

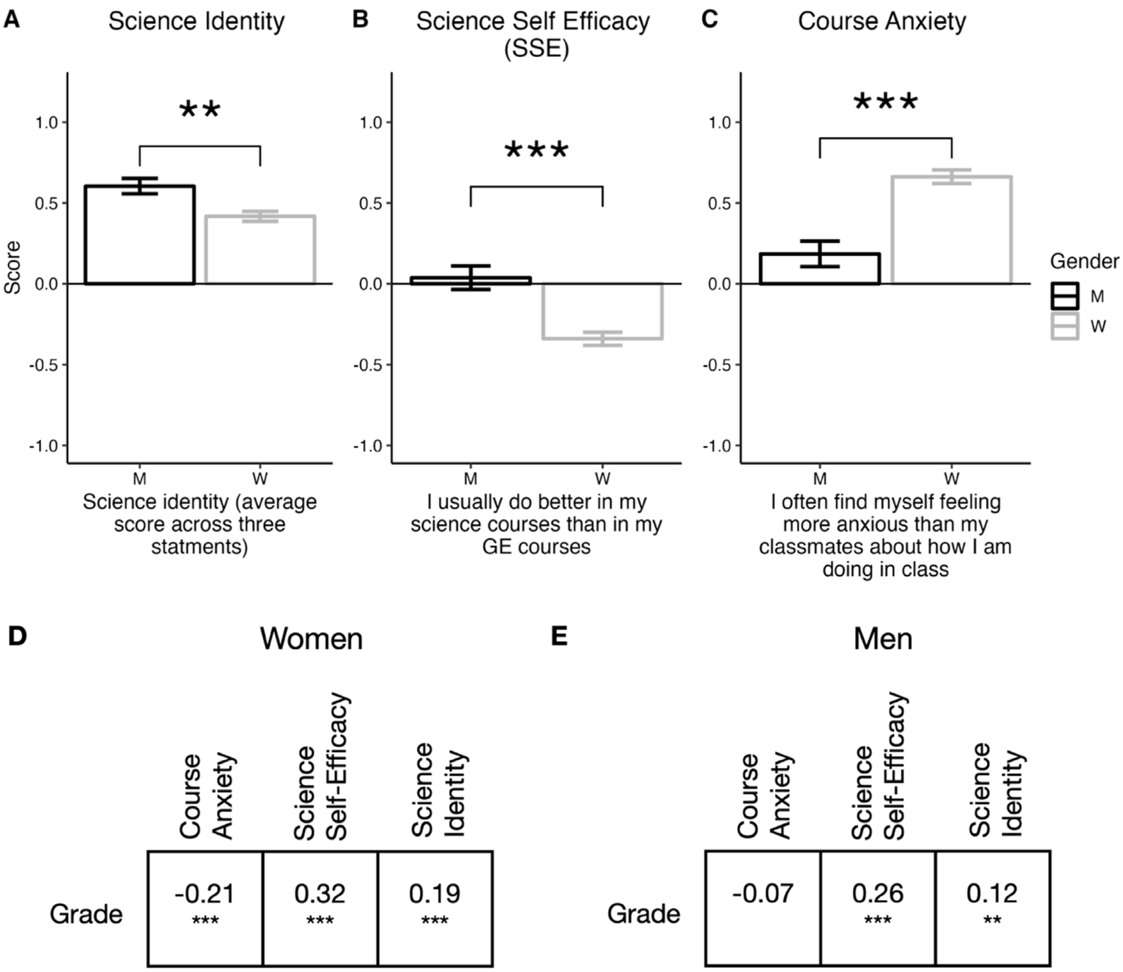


FIGURE 5. Gendered differences in questionnaire responses regarding affective factors are present at the beginning of the course and correlate with course grade outcomes. Women (light gray) agreed less with statements about (A) science identity (averaged across three statements) (two-tailed *t test*: $p < 0.01$) and about (B) science course self-efficacy, but more with statements about (C) course anxiety compared with their male peers (black) (Wilcoxon's rank-sum: $p < 0.001$). In panels A–C, Likert scale scores were centered at 0, where 0 represents "Neutral", positive scores represent agreement with the statement, and negative values represent disagreement. Error bars represent mean and standard error of the mean (SEM). GE = General elective courses. Likert scale scores for science self-efficacy and science identity were significantly correlated with final course grade in (D) women and (E) men, but course anxiety was only significantly correlated with grades in women. Numbers in each box indicate the value of the corresponding correlation (Pearson's correlations). In figure: ** $p < 0.01$; *** $p < 0.001$.

TABLE 7. Robust model estimates for best model for affective factors

Variable	Estimate(β)	Std.Error	p^{a}
(Intercept)	0.030	0.292	0.917
PriorGPA	0.908	0.091	**<0.001**
Gender	–1.178	0.336	**<0.001**
Course Anxiety	0.019	0.037	0.601
Science Self–Efficacy (SSE)	0.058	0.023	**0.013**
Science Identity	0.126	0.031	**<0.001**
PriorGPA * Gender	0.34	0.104	**0.001**
Gender *Course Anxiety	–0.078	0.044	0.077#

[a] *p*-values that were significant ($p < 0.05$) are shown in bold.
$0.05 < p < 0.10$

and science self-efficacy scores ("Agree" for both measures), then the woman who chooses "Strongly Agree" in response to "*I often find myself feeling more anxious than my classmates about how I am doing in my class*" would receive a grade –0.18 points lower than a woman who chooses "Disagree."

DISCUSSION

Using 10 years of student data from an upper-division human physiology course at a large, research-intensive university, we find evidence for a gendered grade disparity that is comparable to gendered grade penalties previously reported in introductory biology courses (Eddy and Brownell, 2016; Odom *et al.*, 2021). These gendered performance gaps were present across a range of systemic advantages conferred by other demographic identities (i.e., PEER status, first generation status, and low socioeconomic status). We also uncovered an interaction between student gender and prior academic performance on course outcomes which moderated the strength of the gender grade penalty but did not eliminate it (even when prior GPA was 4.0). In contrast with studies of lower division biology courses (Bailey *et al.*, 2020; Eddy *et al.*, 2014), the percent of women in the course and the presence of a woman on the instructional team were not associated with student grades. Analysis of student responses to questionnaire revealed that science identity, science self-efficacy, and course anxiety, factors known to be associated with student performance, showed gendered patterns, even at the beginning of the course. Of these affective factors, only the course anxiety score showed a gendered correlation with course grade, with course anxiety having a negative significant correlation with grade for women. This relationship remained even when accounting for prior GPA.

Grade Disparities between Men and Women Differed from Prior GPA Disparities in Nearly Half of Course Offerings

Gendered performance disparities have been extensively documented across a variety of introductory STEM courses, including biology (Eddy *et al.*, 2014; Ballen *et al.*, 2017). Similarly, the limited existing research on upper-division courses also shows gendered disparities (though findings are more heterogeneous; Rauschenberger and Sweeder, 2010; Creech and Sweeder, 2012; Salehi *et al.*, 2019). In our study, we found consistent gendered performance differences across 10 years of offerings of an upper-division biology course, regardless of prior GPA. In addition, we also found that gender showed a significant interaction with prior academic performance, with the gendered grade penalty increasing with decreasing prior GPA. We have not been able to find other studies that reported possible interactions between gender and prior GPA, and thus we cannot speculate as to the prevalence of this interaction across disciplines and course levels.

In addition to an interaction between gender and prior GPA, we also found evidence for a "mismatch" between prior GPA and course grades, in that in nearly half of all course offerings, women entered the course with higher GPA than men but received lower course grades, on average. It is important to note that had we only compared course grades averaged across all offerings and overall prior GPA between men and women (a common approach), we would have missed this mismatch. The mismatch became evident only in examining trends across individual offerings of the course. This difference in approach could account for the lack of similar findings in the literature.

Another measure that can be used to examine the relationship between prior GPA and course grade within individuals is *grade anomaly*. Studies have, in fact, found that grade penalties (i.e., negative grade anomalies) are larger for women than for men in many undergraduate STEM courses (Matz *et al.*, 2017; Malespina and Singh, 2022, 2023). However, grade anomalies cannot discriminate as to whether one group had *equal* or *higher* prior GPAs than the other, as both would result in larger grade anomalies for the group with lower course grades. Our approach of comparing gendered differences in average prior GPA versus gendered differences in course grade allows us to parse whether and when each of the above scenarios occurs by offerings. Uncovering that nearly half of offerings result in a mismatch where women both have higher prior GPAs *and* lower course grades on average emphasizes that this course leads to outcomes that are often opposite and inconsistent with what we would expect based on students' prior experiences and performance at the university. Thus, the view is shifted to trends across course offerings, rather than within individual students. We then can take more of a "course deficit" view (Cotner and Ballen, 2017), examining what elements of the course itself may be disserving students and leading to these inconsistencies with prior experiences in the institution.

Both the mismatches at the level of offering and the significantly different relationship between prior GPA and course grade for gender (the *PriorGPA*Gender* interaction in our linear models) highlight meaningful grade inequities for women in this course. As grade penalties are more common and larger in STEM disciplines (Koester *et al.*, 2016; Matz *et al.*, 2017), and women tend to be more sensitive to grades than men when choosing majors of study (Rask and Tiefenthaler, 2008; Ost, 2010; Ellis *et al.*, 2016; Maries *et al.*, 2022), the contrast between prior GPA and course outcome in an upper-division STEM course could negatively influence women's persistence in STEM majors or careers. Indeed, grade anomalies may be more salient to student decision making than raw course grades (Witteveen and Attewell, 2020; Malespina and Singh, 2023), and women are more likely to change their STEM career plans later in their degree than men (Rosenzweig *et al.*, 2021). Thus, "mismatches" between prior academic performance and upper-division course outcomes such as those described in this study could have significant impacts on individual women's academic decisions and self-concept. Further research is needed to determine whether similar mismatches are observed in other STEM courses and institutions and to examine their possible impact on women's persistence in STEM.

Gendered Differences were Present Across a Variety of Systemic Advantages

Additionally, we found gendered grade disparities across the spectrum of systemic advantages (as conferred by race/ethnicity, socioeconomic status, and first-generation status). The presence of gendered grade differences in students both with few and many systemic advantages, and the lack of significant interactions between gender and other demographic factors suggest that gender may relate to course outcomes in a different manner than these other axes of identity. In addition, PEER status, low socioeconomic status and first-generation status exhibited more "linear" relationships between prior GPA and course grade in this course (i.e., these demographic factors did not significantly interact with prior GPA in linear models). While prior GPA "explained away" a significant amount of the disparities in performance across socioeconomic status, PEER, and first-generation status, it had minimal impact on the observed gender disparities. These results are similar to those seen in introductory chemistry courses, where raw and academic performance-adjusted gaps significantly differ for PEER, low socioeconomic status students and first-generation students, but not for women (Harris *et al.*, 2020). Our results imply that students with these identities have experiences in an upper-division biology course (physiology) that are consistent with systemic inequities being experienced across all undergraduate coursework, a pattern that must be examined further. Below, we discuss what student and course level factors–academic, environmental, or affective–may be contributing to the persistent gender gap observed in the upper-division biology course under study.

Neither Student Majors, Instructor Gender, Nor Gender Representation in the Class were Related to Gender Gaps in the Course

We hypothesized that the observed gender gap in upper-division physiology course could be associated with gendered differences in study major and/or prior STEM preparation. If men and women differ in their declared majors, as prior evidence suggests (Griffith, 2010; Shapiro and Sax, 2011), then the observed mismatches between prior GPA and course grades for women could be explained by differences in their academic preparation for the course (related to their own major required courses) and the differences in the grading practices across disciplines (which would impact their prior GPA; Nord *et al.*, 2011). However, the gendered "mismatch" between prior GPA and course grade was present across the top six majors present in the course, regardless of college. Furthermore, men and women did not significantly differ in their prior STEM units taken, and prior STEM units did not correlate with course grades. These results are consistent with other studies of upper-division biology courses, where major of study did not have a significant effect on grade (Rauschenberger and Sweeder, 2010; Creech and Sweeder, 2012). Major of study may have less impact at the upper-division level, where almost all students taking STEM courses are STEM majors. Indeed, this explanation was offered as a reason for why no gender gaps were observed in upper-division biology and engineering courses in another study (Salehi *et al.*, 2019). However, the upper-division biology course examined in this study had students from over 100 majors, as it is a common prerequisite for a variety of health sciences programs. Differences in declared major, at least within the life and biomedical science majors we examined, did not seem to be associated with the observed gender grade disparities.

We also did not find evidence that the representation of women in the student body or instructional team were associated with student grades in this course. This result contrasts with previous studies. Environmental factors, such as gender composition of the course or the presence of a woman on the instructional team has been associated with improved course outcomes for women in STEM courses (Bowman *et al.*, 2022), including in biology (Eddy *et al.*, 2014; Bailey *et al.*, 2020). There are a few possible reasons why we did not find this relationship in our course. First, the range of representation of women, in both the student body and instructional team, was more limited in our study than others. For instance, Bailey *et al.* (2020) examined courses where the percent of women attending ranged from 20 to 80%, but our class offerings ranged from 59 to 74% women. Similarly, studies that found significant interactions between women's grades and instructor gender did so in courses taught *solely* by women (Lauer *et al.*, 2013; Eddy *et al.*, 2014; Bailey *et al.*, 2020); our course was typically co–taught and included at least one man; only two offerings had women-only instructional teams. Thus, the effect of gender representation may be diluted when a course is always majority women, and the instructional team is mixed gender. Second, instructional choices that may vary with instructor gender, such as assessment and grading, were held constant in this course (as per departmental policy). For instance, multiple-choice, timed exams were exclusively used regardless of the individual instructor, removing the possibility of subjective grading or assessment differences. Last, the large overall class size, rather than gender makeup of the student body, may have more of an effect on women (Ballen *et al.*, 2018; Odom *et al.*, 2021), especially in a large class that is always majority-women.

We did not have records of daily attendance, nor did we measure how students may have perceived the gender composition or size of the course in this study. We also were not able to find studies describing how class size, student gender, or other environmental factors are associated with student perceptions of these variables. Comparisons of the relationship between student perceptions of overall class size and gender classroom representation and the actual classroom composition would help illuminate how these environmental factors may influence student outcomes, affect, and attitudes.

Gendered Differences in Course Anxiety may Contribute to Gendered Grade Disparities in Upper-Division Biology Courses

Reported science self-efficacy and science identity at the beginning of the upper-division biology course under study were correlated with final course grade for men and women. In contrast, reported course anxiety was only correlated with course outcomes for women. This result aligns with a growing body of research showing that test anxiety may influence exam performance for women, but not men, in introductory (Ballen *et al.*, 2017; Lowe, 2019; Cotner *et al.*, 2020; Odom *et al.*, 2021) and upper-division biology (Salehi *et al.*, 2019). We also found that the effect of course anxiety on women's grades persisted even when prior academic preparation was accounted for using multilevel modeling, suggesting that anxiety may contribute to some of the misalignment between prior academic performance and grades seen for women in our course.

Gendered differences in course anxiety may disproportionately affect women when combined with traditional, high-stakes exam assessment. Our human physiology course used exclusively multiple-choice, timed exams to assess student learning, and this may have interacted with course anxiety differences to shape the gender performance gap. High-stakes exams, combined with traditional lecture formats, are associated with gendered performance gaps across biology courses (Odom *et al.*, 2021). When courses incorporate multiple forms of assessment, especially low stakes assignments like projects, quizzes, and homework, gender gaps can be reduced or even reversed (Stanger-Hall, 2012; Ballen *et al.*, 2017; Cotner and Ballen, 2017; Salehi *et al.*, 2019). Indeed, differences in assessment structures has been hypothesized as a reason why gender gaps were not observed in some upper-division biology courses, as compared with larger introductory biology (Salehi *et al.*, 2019).

Importantly, these differences in course anxiety associated with gender likely reflect institutionalized experiences of women across STEM education, rather than intrinsic gender differences in propensity towards anxiety. For instance, women may experience sex discrimination, be it overtly or through microaggressions (Sue, 2010; Sekaquaptewa, 2019) in STEM contexts (Steele *et al.*, 2002; Funk and Parker, 2018). These experiences may lead to increased stereotype threat, where women are exposed to attitudes that convey the message that women underperform or are "inferior" in scientific ability, and thus, feel psychological pressure that their own performance will reinforce this negative stereotype (Spencer *et al.*, 1999; Schmader, 2002). Indeed, stereotype threat has been proposed to underlie gendered differences in test anxiety (Spencer *et al.*, 1999; Osborne, 2001), and can lead women to leave STEM fields (Steele *et al.*, 2002). Additionally, women may experience imposter syndrome (Clance and Imes, 1978) where they feel inadequate, and a lack of belonging in science, despite their own performance as evidence to the contrary. These systemic and institutionalized experiences may accumulate as women progress through their STEM degree (Cromley *et al.*, 2013), leading to women perceiving that they feel more anxious about course outcomes compared with their peers. Future work is still needed to address what experiences may underlie this sense of relative anxiety, and what aspects of those experiences may relate to students' performance in high-stakes settings. However, our study shows that gendered differences in course anxiety persist into the upper-division level, and these differences may possibly result from institutionalized experiences of gender stereotypes in STEM education.

Constraints of Final Grade Data

As we only had access to final registrar grades, we could not evaluate the effects of variation in instructor grading practices on student outcomes in the course in this study. While high-stakes, multiple-choice exams were consistently used across all offerings, instructors may have differed in the "cut offs" or heuristics used to assign letter grades. This course is traditionally graded "on a curve" a practice that researchers argue is not objective and inconsistent and often promotes competition over learning (Bowen and Cooper, 2022). Further, grade assignment may be subject to gender bias. Instructors may subconsciously view men more favorably than women during grade assignment as has been observed in studies of implicit bias when evaluating graduate school applicants (Moss-Racusin *et al.*, 2012). Gender may also affect the degree of student involvement in the grading process; men tend to be more "academically entitled" (Ciani *et al.*, 2008) and are more likely to request regrades on assignments than women (Li and Zafar, 2020). Grade assignment practices remain an understudied area in biology education research, and more work is needed to understand how these practices affect student outcomes.

Additionally, without individual exam scores, we could not examine how gender gaps may have emerged at the exam level or whether they were consistent throughout all course exams. Exams may have varied across offerings in multiple ways that could have shaped the gender gap. First, variation in relative exam weight, or the percent of the total course grade each exam constitutes, could have led to differential gender gaps. When manipulated within a single class, high weight exams showed larger gender gaps than exams where the weight was lower (e.g., exams that could be dropped), where the gender gap actually reversed (Montolio and Taberner, 2021). This pattern aligns with studies showing that women outperform men on lower-stakes assignments (e.g., quizzes, homework, projects) compared with exams (Ballen *et al.*, 2017; Cotner and Ballen, 2017; Salehi *et al.*, 2019). Second, variation in exam question difficulty may contribute to gendered performance gaps as one study has shown that women underperform on higher Bloom's taxonomy exam questions in introductory biology (Wright *et al.*, 2016), though more research remains to be done. Last, limited time per exam question may have disproportionately affected women. All exams were limited to the 50-minute class period, but exam length, and thus time per question, may have varied across offerings. When exams are not timed (Miller *et al.*, 1994) or when the length of the exam/cognitive test is longer (Balart and Oosterveen, 2019), gender gaps can be reduced (but see evidence to the contrary in physics; Tarchinski *et al.*, 2022). Importantly, all these studies highlight the significant impact of course structure, curriculum choices and pedagogical practices on student outcomes. While the mechanisms by which these changes impact gender outcome disparities remain understudied, these changes likely interact with affective factors such as test anxiety and stereotype threat, which themselves are likely to be the result of institutionalized biases and are not due to differences in objective ability (Spencer *et al.*, 1999). Future studies in upper-division biology courses with access to exam-level data should evaluate the impact of assessment changes on gendered grade disparities and affective factors like anxiety.

Limitations of the Study

As with final grades, we were limited to registrar data for demographic information, much of which is collected in a binary manner. For instance, the impact of course structures on students whose identities do not fit these binary definitions, such as those of transgender or gender nonconforming students (Cooper *et al.*, 2020) and students with other concealable, stigmatized identities (Chaudoir and Quinn, 2010) cannot be examined with registrar data alone. While we incorporated affective factors from course questionnaire data, this questionnaire data was limited to only three offerings of the course and may not encompass the trends seen over the full ten-year period. Further, we were limited to a single item to evaluate science self-efficacy and course anxiety, rather than using validated scales. We were therefore unable to perform exploratory factor analysis to validate these tools as they are only single items. Our questionnaire also addressed "course anxiety" rather than "test anxiety" per se. Despite these limitations, we still found a significant gendered difference in course anxiety that is consistent with published literature (Ballen *et al.*, 2017; Cotner and Ballen, 2017; Salehi *et al.*, 2019). Future studies of these affective factors should validate and use multi-item scales to assess self-efficacy and anxiety in upper-division contexts. Last, the trends we observed in this single course until 2019 may have changed significantly after the impact of the COVID-19 pandemic and use of online and hybrid environments (e.g. Mohammed *et al.*, 2021; Ewell *et al.*, 2022), and more recent trends should be compared in future studies. Below, we highlight opportunities for future research to expand upon our study and further address these gendered grade disparities.

Recommendations for Future Research

The gendered performance gaps we observed could be addressed by course modifications, ranging from those that could be implemented by individual instructors, to the addition of supplemental instruction sections at the departmental level. While there is a growing body of research on interventions at the lower division level (e.g., Freeman *et al.*, 2014; Wu *et al.*, 2021) which can be used to address gendered grade disparities (Karim *et al.*, 2018; Harris *et al.*, 2019), few studies have evaluated these strategies in upper-division courses. Importantly, many of these interventions and modifications can benefit all students in the course, not just women, making courses more equitable overall.

Given that we found gendered differences in anxiety, an ideal starting point may be interventions that directly address student course anxiety. Classroom interventions designed to ameliorate test anxiety and/or stereotype threat show some promise for closing gendered performance gaps in introductory STEM courses, though results can vary (Miyake *et al.*, 2010; Jordt *et al.*, 2017; Harris *et al.*, 2019). Such interventions can include exercises that encourage: 1) emotional reappraisal, where students read information that frames physiological arousal associated with stress as beneficial for performance (Jamieson *et al.*, 2010, 2013; Harris *et al.*, 2019), 2) expressive writing, which aims to ameliorate student anxieties by removing worries from working memory before exams (Frattaroli *et al.*, 2011; Ramirez and Beilock, 2011), and 3) values affirmation, where students write and reflect on aspects of their lives they consider valuable to combat stereotype threat (Wu *et al.*, 2021). These exercises could be relatively straightforward for individual instructors to implement in a single class period (Yeager and Walton, 2011; Borman, 2017), offering a feasible starting place for making courses more equitable.

Instructors could also change course assessment and instructional methods with the goal of improving course outcomes for all students. Research shows that incorporating mixed assessments beyond exams may decrease test anxiety and benefit women in introductory biology courses (Ballen *et al.*, 2017; Cotner and Ballen, 2017; Salehi *et al.*, 2019). Decreasing high-stakes nature of exams, such as changing the percent of the total grade comprised by exams, increasing exam time, or removing time limits on exams, as well as diversifying question types, could also reduce outcome disparities (Weaver and Raptis, 2001; Wright *et al.*, 2016; Montolio and Taberner, 2021). Instructors can also employ active learning in the large lecture setting, an approach that has been shown to improve student performance (Theobald *et al.*, 2020). However, recent work shows that some active learning techniques can heighten student anxiety (England *et al.*, 2017; Cooper *et al.*, 2018; Downing *et al.*, 2020), in ways that can disproportionately affect women (Eddy *et al.*, 2015; Aguillon *et al.*, 2020). Such outcomes would be counterproductive to reducing gendered differences in anxiety, and thus, instructors should carefully consider the impacts of any strategies they utilize. Future research should investigate the respective impacts of classroom structure and assessment formats on test anxiety and gendered equity gaps at the upper-division level.

Last, at the departmental level, the addition of supplemental instruction (SI) sections holds promise for improving equity in large enrollment upper-division courses such as the one in this study. Supplemental instruction (SI) allows students to engage with difficult course material in smaller class sections that center collaborative learning and peer-engagement (Martin and Arendale, 1992). Enrollment in SI in introductory STEM courses has been shown to particularly benefit students from racially marginalized backgrounds (Rath *et al.*, 2007; Peterfreund *et al.*, 2008; Anfuso *et al.*, 2022) and women (Rabitoy *et al.*, 2015; Shapiro *et al.*, 2016; Cole *et al.*, 2018). SI sections may yield these benefits because they incorporate many of the previously mentioned strategies that address anxiety and other affective factors; they employ active learning techniques in smaller classroom settings, use lower-stake assessments (e.g., “pass/no-pass” grading; Rath *et al.*, 2007), and involve peer and instructor interactions that may promote student sense of belonging and perceived representation. While SI sections are common practice in lower division courses, they are rarely implemented for upper-division courses, and as a result, little is known as to their possible impact at this course level. Given the benefits seen in lower division courses (Martin and Arendale, 1992; Dawson *et al.*, 2014), adding SI sections to large enrollment upper-division courses could likely ameliorate some of the outcome disparities seen across demographic factors in this study.

Importantly, future analyses should expand our work to examine if the gaps we observed in this single course are present in other upper-division biology courses and across institutions (Thompson *et al.*, 2020). While this study focuses on only one course at a single institution, it echoes prior patterns seen in other upper-division biomedical courses (Rauschenberger and Sweeder, 2010; Creech and Sweeder, 2012). While upper-division STEM courses are not usually considered “gateway” courses, many have large enrollments and are part of a “core” sequence required for graduation in a major. The initial courses in such sequences are uniquely placed to become “upper-division gateways” that could still affect student retention, graduation rates and opportunities for postgraduate careers. Future work should examine the prevalence of equity disparities in student outcomes in upper-division STEM courses across disciplines and institutions and how these may contribute to late-stage student attrition. While such multi-institutional studies have been conducted for introductory STEM courses (Matz *et al.*, 2017; Castle, 2021), an examination of upper-division coursework is still needed. The presence of similar outcome disparities in several institutions would point to structural issues in upper-division courses across the higher education system that have hitherto been hidden. Ideally, such research will guide systemic and institutional change that can improve outcomes for women *throughout* the course of the undergraduate degree and promote retention of women in STEM nationwide.

CONCLUSIONS

We found evidence of gendered equity gaps in course grades in a large-enrollment upper-division biology course across 10 years of data and 35 offerings. These gendered grade disparities persisted even when other academic and demographic factors were accounted for in multilevel models and did not appear to relate to environmental factors like course size, gender ratio of the classroom, and instructor gender. Detailed analysis of the relationship between course grades and prior GPA revealed a gendered preparation-outcome mismatch such that men received higher grades even in offerings when women had stronger academic preparation. We also identified gendered differences in self-efficacy, science identity and course anxiety that were associated to student outcomes in the upper division course. Together, our results underscore the importance of examining gendered equity gaps and outcome-related psychosocial factors beyond the introductory level, and a growing need to apply pedagogical interventions to close such gaps in upper-division courses.

ACKNOWLEDGMENTS

We thank Bradley Velasquez and Meryl Motika at the UC Davis Center for Educational Effectiveness for their invaluable help with data access and processing. We thank Robert Furrow, Caitlin Hawley, and Rebecca Matz for helpful conversations and comments, as well as two anonymous reviewers who provided feedback that greatly improved the manuscript. We also wish to acknowledge the members of the Sloan Equity and Inclusion in STEM Introductory Courses (SEISMIC; www.seismicproject.org) collaboration Measurements Working Group for insights and their efforts to develop novel approaches for examining equity gaps in STEM courses.

REFERENCES

AAMC. (2018). *Diversity in medicine: facts and figures 2018*. Association of American Medical Colleges. Retrieved October 1, 2022, from https://www.aamc.org/data-reports/workforce/interactive-data/figure-19-percentage-physicians-sex-2018

AAMC. (2019). *2019 Fall applicant, matriculant, and enrollment data tables* (pp. 15). Association of American Medical Colleges. Retrieved October 1, 2022, from https://www.aamc.org/system/files/2019-12/2019%20AAMC%20Fall%20Applicant%2C%20Matriculant%2C%20and%20Enrollment%20Data%20Tables_0.pdf

AggieDash. (2022). *AggieDash*. Retrieved November 23, 2022, from https://aggiedash.ucdavis.edu/#/views/UndergraduateCohortRetentionand-GraduationRates_0/RetentionandGraduation?:iId=3

Aguillon, S. M., Siegmund, G.-F., Petipas, R. H., Drake, A. G., Cotner, S., & Ballen, C. J. (2020). Gender differences in student participation in an active-learning classroom. *CBE—Life Sciences Education*, *19*(2), ar12. https://doi.org/10.1187/cbe.19-03-0048

Ahn, T., Arcidiacono, P., Hopson, A., & Thomas, J. (2019). Equilibrium grade inflation with implications for female interest in STEM majors (No. w26556; p. w26556). *National Bureau of Economic Research*, https://doi.org/10.3386/w26556

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Anfuso, C., Awong-Taylor, J., Curry Savage, J., Johnson, C., Leader, T., Pinzon, K., … & Achat-Mendes, C. (2022). Investigating the impact of peer supplemental instruction on underprepared and historically underserved students in introductory STEM courses. *International Journal of STEM Education*, *9*(1), 55. https://doi.org/10.1186/s40594-022-00372-w

Appel, M., Kronberger, N., & Aronson, J. (2011). Stereotype threat impairs ability building: effects on test preparation among women in science and technology: stereotype threat and ability building. *European Journal of Social Psychology*, *41*(7), 904–913. https://doi.org/10.1002/ejsp.835

Asai, D. (2020). Excluded. *Journal of Microbiology & Biology Education*, *21*(1), 754–755. https://doi.org/10.1128/jmbe.v21i1.2071

Aulck, L., & West, J. (2017). Attrition and performance of community college transfers. *PLOS ONE*, *12*(4), e0174683. https://doi.org/10.1371/journal.pone.0174683

Bailey, E. G., Greenall, R. F., Baek, D. M., Morris, C., Nelson, N., Quirante, T. M., … & Williams, K. R. (2020). Female in-class participation and performance increase with more female peers and/or a female instructor in life sciences courses. *CBE—Life Sciences Education*, *19*(3), ar30. https://doi.org/10.1187/cbe.19-12-0266

Balart, P., & Oosterveen, M. (2019). Females show more sustained performance during test-taking than males. *Nature Communications*, *10*(1), 3798. https://doi.org/10.1038/s41467-019-11691-y

Ballen, C. J., Aguillon, S. M., Awwad, A., Bjune, A. E., Challou, D., Drake, A. G., … & Cotner, S. (2019). Smaller classes promote equitable student participation in STEM. *BioScience*, *69*(8), 669–680. https://doi.org/10.1093/biosci/biz069

Ballen, C. J., Aguillon, S. M., Brunelli, R., Drake, A. G., Wassenberg, D., Weiss, S. L., … & Cotner, S. (2018). Do small classes in higher education reduce performance gaps in STEM? *BioScience*, *68*(8), 593–600. https://doi.org/10.1093/biosci/biy056

Ballen, C. J., Salehi, S., & Cotner, S. (2017). Exams disadvantage women in introductory biology. *PLOS ONE*, *12*(10), e0186419. https://doi.org/10.1371/journal.pone.0186419

Barton, K. (2020). *MuMIn: multi-model inference* [R package version 1.43.47].

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bettencourt, G. M., Manly, C. A., Kimball, E., & Wells, R. S. (2020). STEM degree completion and first-generation college students: a cumulative disadvantage approach to the outcomes gap. *The Review of Higher Education*, *43*(3), 753–779. https://doi.org/10.1353/rhe.2020.0006

Bloodhart, B., Balgopal, M. M., Casper, A. M. A., Sample McMeeking, L. B., & Fischer, E. V. (2020). Outperforming yet undervalued: Undergraduate women in STEM. *PLOS ONE*, *15*(6), e0234685. https://doi.org/10.1371/journal.pone.0234685

Bolker, B. (2022). *bbmle: Tools for general maximum likelihood estimation (1.0.25)*. Retrieved February 3, 2023, from https://cran.r-project.org/web/packages/bbmle/index.html

Borman, G. D. (2017). Advancing values affirmation as a scalable strategy for mitigating identity threats and narrowing national achievement gaps. *Proceedings of the National Academy of Sciences*, *114*(29), 7486–7488. https://doi.org/10.1073/pnas.1708813114

Bourke, B. (2016). Meaning and implications of being labelled a predominantly white institution. *College and University*, *91*(3), 12–18.

Bowen, R. S., & Cooper, M. M. (2022). Grading on a curve as a systemic issue of equity in chemistry education. *Journal of Chemical Education*, *99*(1), 185–194. https://doi.org/10.1021/acs.jchemed.1c00369

Bowman, N. A., Logel, C., LaCosse, J., Jarratt, L., Canning, E. A., Emerson, K. T. U., & Murphy, M. C. (2022). Gender representation and academic achievement among STEM-interested students in college STEM courses. *Journal of Research in Science Teaching*, *59*(10):1876-1900. https://doi.org/10.1002/tea.21778

Burnham, K. P., & Anderson, D. R. (Eds.) (2004). *Model selection and multimodel inference*. New York, NY: Springer. https://doi.org/10.1007/b97636

Carrell, S. E., Page, M. E., & West, J. E. (2010). Sex and science: how professor gender perpetuates the gender gap. *Quarterly Journal of Economics*, *125*(3), 1101–1144. https://doi.org/10.1162/qjec.2010.125.3.1101

Casper, A. M. A., Rebolledo, N., Lane, A. K., Jude, L., & Eddy, S. L. (2022). "It's completely erasure": a qualitative exploration of experiences of transgender, nonbinary, gender nonconforming, and questioning students in biology courses. *CBE—Life Sciences Education*, *21*(4), ar69. https://doi.org/10.1187/cbe.21-12-0343

Castle, S. (2021). Equity in the STEM landscape: a multi-institutional approach to mapping systemic advantages within STEM courses. *Proceedings of the 2021 AERA Annual Meeting*. 2021 AERA Annual Meeting. https://doi.org/10.3102/1689325

Chang, M. J., Sharkness, J., Hurtado, S., & Newman, C. B. (2014). What matters in college for retaining aspiring scientists and engineers from underrepresented racial groups: retaining aspiring scientists. *Journal of Research in Science Teaching*, *51*(5), 555–580. https://doi.org/10.1002/tea.21146

Chaudhary, A. M. D., Naveed, S., Safdar, B., Saboor, S., Zeshan, M., & Khosa, F. (2021). Gender differences in research project grants and R01 grants at the National Institutes of Health. *Cureus*, *13*(5):e14930 https://doi.org/10.7759/cureus.14930

Chaudoir, S. R., & Quinn, D. M. (2010). Revealing concealable stigmatized identities: The impact of disclosure motivations and positive first-disclosure experiences on fear of disclosure and well-being: revealing concealable stigmatized identities. *Journal of Social Issues*, *66*(3), 570–584. https://doi.org/10.1111/j.1540-4560.2010.01663.x

Ciani, K. D., Summers, J. J., & Easter, M. A. (2008). Gender differences in academic entitlement among college students. *The Journal of Genetic Psychology*, *169*(4), 332–344. https://doi.org/10.3200/GNTP.169.4.332-344

Cimpian, J. R., Kim, T. H., & McDermott, Z. T. (2020). Understanding persistent gender gaps in STEM. *Science*, *368*(6497), 1317–1319. https://doi.org/10.1126/science.aba7377

Clance, P. R., & Imes, S. A. (1978). The imposter phenomenon in high achieving women: dynamics and therapeutic intervention. *Psychotherapy: Theory, Research & Practice*, *15*(3), 241–247. https://doi.org/10.1037/h0086006

Cole, T., Kaeli, E., Priem, B., Ghio, C., DiMilla, P., & Reisberg, R. (2018). The influence of preconceptions, experience, and gender on use of supplemental instruction and academic success in a freshman chemistry course for engineers. *2018 ASEE Annual Conference & Exposition Proceedings*. https://doi.org/10.18260/1-2–31116

Cooper, K. M., Auerbach, A. J. J., Bader, J. D., Beadles-Bohling, A. S., Brashears, J. A., Cline, E., … & Brownell, S. E. (2020). Fourteen recommendations to create a more inclusive environment for LGBTQ+ individuals in academic biology. *CBE—Life Sciences Education*, *19*(3), es6. https://doi.org/10.1187/cbe.20-04-0062

Cooper, K. M., Downing, V. R., & Brownell, S. E. (2018). The influence of active learning practices on student anxiety in large-enrollment college science classrooms. *International Journal of STEM Education*, *5*(1), 23. https://doi.org/10.1186/s40594-018-0123-6

Cooper, K. M., Eddy, S. L., & Brownell, S. E. (2023). Research anxiety predicts undergraduates' intentions to pursue scientific research careers. *CBE—Life Sciences Education*, *22*(1), ar11. https://doi.org/10.1187/cbe.22-02-0022

Cooper, K. M., Gin, L. E., Akeeh, B., Clark, C. E., Hunter, J. S., Roderick, T. B., … & Brownell, S. E. (2019). Factors that predict life sciences student persistence in undergraduate research experiences. *PLOS ONE*, *14*(8), e0220186. https://doi.org/10.1371/journal.pone.0220186

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

Cotner, S., Jeno, L. M., Walker, J. D., Jørgensen, C., & Vandvik, V. (2020). Gender gaps in the performance of Norwegian biology students: The roles of test anxiety and science confidence. *International Journal of STEM Education*, *7*(1), 55. https://doi.org/10.1186/s40594-020-00252-1

Creech, L. R., & Sweeder, R. D. (2012). Analysis of student performance in large-enrollment life science courses. *CBE—Life Sciences Education*, *11*(4), 386–391. https://doi.org/10.1187/cbe.12-02-0019

Crenshaw, K. (1991). Mapping the margins: intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, *43*(6), 1241. https://doi.org/10.2307/1229039

Cromley, J. G., Perez, T., Wills, T. W., Tanaka, J. C., Horvat, E. M., & Agbenyega, E. T.-B. (2013). Changes in race and sex stereotype threat among diverse STEM students: Relation to grades and retention in the majors. *Contemporary Educational Psychology*, *38*(3), 247–258. https://doi.org/10.1016/j.cedpsych.2013.04.003

Dawson, P., van der Meer, J., Skalicky, J., & Cowley, K. (2014). On the effectiveness of supplemental instruction: a systematic review of supplemental instruction and peer-assisted study sessions literature between 2001 and 2010. *Review of Educational Research*, *84*(4), 609–639. https://doi.org/10.3102/0034654314540007

De Brey, C., Snyder, T. D., Zhang, A., & Dillow, S. A. (2021). *Digest of Education Statistics 2019* (pp. 651). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Retrieved March 5, 2022, from https://nces.ed.gov/pubs2021/2021009.pdf

Downing, V. R., Cooper, K. M., Cala, J. M., Gin, L. E., & Brownell, S. E. (2020). Fear of negative evaluation and student anxiety in community college active-learning science courses. *CBE—Life Sciences Education*, *19*(2), ar20. https://doi.org/10.1187/cbe.19-09-0186

Dziak, J. J., Coffman, D. L., Lanza, S. T., & Li, R. (2012). *Sensitivity and specificity of information criteria* (Technical Report No. 12–119). University Park, PA: The Methodology Center, The Pennsylvania State University. Retrieved November 12, 2021, from https://www.methodology.psu.edu/files/2019/03/12-119-2e90hc6.pdf

Eagan, M. K., Hurtado, S., & Chang, M. J. (2010). *What matters in STEM: Institutional contexts that influence STEM bachelor's degree completion rates*. Retrieved November 12, 2021, from https://heri.ucla.edu/nih/downloads/ASHE%202010%20-%20Eagan%20Hurtado%20Chang%20-%20STEM%20Completion.pdf

Eaton, A. A., Saunders, J. F., Jacobson, R. K., & West, K. (2020). How gender and race stereotypes impact the advancement of scholars in STEM: professors' biased evaluations of physics and biology post-doctoral candidates. *Sex Roles*, *82*(3–4), 127–141. https://doi.org/10.1007/s11199-019-01052-w

Eccles, J. S., & Wigfield, A. (2020). From expectancy-value theory to situated expectancy-value theory: a developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary Educational Psychology*, *61*, 101859. https://doi.org/10.1016/j.cedpsych.2020.101859

Eddy, S. L., & Brownell, S. E. (2016). Beneath the numbers: a review of gender disparities in undergraduate education across science, technology, engineering, and math disciplines. *Physical Review Physics Education Research*, *12*(2), 020106. https://doi.org/10.1103/Physrevphysedures.12.020106

Eddy, S. L., Brownell, S. E., Thummaphan, P., Lan, M.-C., & Wenderoth, M. P. (2015). Caution, student experience may vary: social identities impact a student's experience in peer discussions. *CBE—Life Sciences Education*, *14*(4), ar45. https://doi.org/10.1187/cbe.15-05-0108

Eddy, S. L., Brownell, S. E., & Wenderoth, M. P. (2014). Gender gaps in achievement and participation in multiple introductory biology classrooms. *CBE—Life Sciences Education*, *13*(3), 478–492. https://doi.org/10.1187/cbe.13-10-0204

Ellis, J., Fosdick, B. K., & Rasmussen, C. (2016). Women 1.5 times more likely to leave STEM pipeline after calculus compared to men: lack of mathematical confidence a potential culprit. *PLOS ONE*, *11*(7), e0157447. https://doi.org/10.1371/journal.pone.0157447

England, B. J., Brigati, J. R., & Schussler, E. E. (2017). Student anxiety in introductory biology classrooms: perceptions about active learning and persistence in the major. *PLOS ONE*, *12*(8), e0182506. https://doi.org/10.1371/journal.pone.0182506

Ewell, S. N., Josefson, C. C., & Ballen, C. J. (2022). Why did students report lower test anxiety during the COVID-19 pandemic? *Journal of Microbiology & Biology Education*, *23*(1), e00282–21. https://doi.org/10.1128/jmbe.00282-21

Foraker, M. J. (2012). *Does changing majors really affect the time to graduate? The impact of changing majors on student retention, graduation, and time to graduate*. Bowling Green, KY: Western Kentucky University. Retrieved March 1, 2020, from https://www.semanticscholar.org/paper/Does-Changing-Majors-Really-Affect-the-Time-to-The-Foraker/cb8df7853c6937092ec842fdc9f674b5a4767f68

Fox, C. W., & Paine, C. E. T. (2019). Gender differences in peer review outcomes and manuscript impact at six journals of ecology and evolution. *Ecology and Evolution*, *9*(6), 3599–3619. https://doi.org/10.1002/ece3.4993

Frattaroli, J., Thomas, M., & Lyubomirsky, S. (2011). Opening up in the classroom: effects of expressive writing on graduate school entrance exam performance. *Emotion*, *11*(3), 691–696. https://doi.org/10.1037/a0022946

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, *111*(23), 8410–8415. https://doi.org/10.1073/pnas.1319030111

Friedmann, E., & Efrat-Treister, D. (2023). Gender bias in stem hiring: implicit in-group gender favoritism among men managers. *Gender & Society*, *37*(1), 32–64. https://doi.org/10.1177/08912432221137910

Funk, C., & Parker, K. (2018). *Women and men in STEM often at odds over workplace equity*. Washington, DC: Pew Research Center. Retrieved March 1, 2022, from https://assets.pewresearch.org/wp-content/uploads/sites/3/2018/01/09142305/PS_2018.01.09_STEM_FINAL.pdf

Garvey, J. C., & Rankin, S. R. (2015). The influence of campus experiences on the level of outness among trans-spectrum and queer-spectrum students. *Journal of Homosexuality*, *62*(3), 374–393. https://doi.org/10.1080/00918369.2014.977113

Griffith, A. L. (2010). Persistence of women and minorities in STEM field majors: Is it the school that matters? *Economics of Education Review*, *29*(6), 911–922. https://doi.org/10.1016/j.econedurev.2010.06.010

Grunspan, D. Z., Eddy, S. L., Brownell, S. E., Wiggins, B. L., Crowe, A. J., & Goodreau, S. M. (2016). Males under-estimate academic performance of their female peers in undergraduate biology classrooms. *PLOS ONE*, *11*(2), e0148405. https://doi.org/10.1371/journal.pone.0148405

Harris, R. B., Grunspan, D. Z., Pelch, M. A., Fernandes, G., Ramirez, G., & Freeman, S. (2019). Can test anxiety interventions alleviate a gender gap in an undergraduate STEM course? *CBE—Life Sciences Education*, *18*(3), ar35. https://doi.org/10.1187/cbe.18-05-0083

Harris, R. B., Mack, M. R., Bryant, J., Theobald, E. J., & Freeman, S. (2020). Reducing achievement gaps in undergraduate general chemistry could lift underrepresented students into a "hyperpersistent zone." *Science Advances*, *6*(24), eaaz5687. https://doi.org/10.1126/sciadv.aaz5687

Hechtman, L. A., Moore, N. P., Schulkey, C. E., Miklos, A. C., Calcagno, A. M., Aragon, R., & Greenberg, J. H. (2018). NIH funding longevity by gender. *Proceedings of the National Academy of Sciences*, *115*(31), 7943–7948. https://doi.org/10.1073/pnas.1800615115

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Jamieson, J. P., Mendes, W. B., Blackstock, E., & Schmader, T. (2010). Turning the knots in your stomach into bows: reappraising arousal improves performance on the GRE. *Journal of Experimental Social Psychology*, *46*(1), 208–212. https://doi.org/10.1016/j.jesp.2009.08.015

Jamieson, J. P., Mendes, W. B., & Nock, M. K. (2013). Improving acute stress responses: the power of reappraisal. *Current Directions in Psychological Science*, *22*(1), 51–56. https://doi.org/10.1177/0963721412461500

Jordt, H., Eddy, S. L., Brazil, R., Lau, I., Mann, C., Brownell, S. E., ... & Freeman, S. (2017). Values affirmation intervention reduces achievement gap between underrepresented minority and white students in introductory biology classes. *CBE—Life Sciences Education*, *16*(3), ar41. https://doi.org/10.1187/cbe.16-12-0351

Kanny, M. A., Sax, L. J., & Riggers-Piehl, T. A. (2014). Investigating forty years of stem research: how explanations for the gender gap have evolved over time. *Journal of Women and Minorities in Science and Engineering*, *20*(2), 127–148. https://doi.org/10.1615/JWomenMinorScienEng.2014007246

Karim, N. I., Maries, A., & Singh, C. (2018). Do evidence-based active-engagement courses reduce the gender gap in introductory physics? *European Journal of Physics*, *39*(2), 025701. https://doi.org/10.1088/1361-6404/aa9689

Koch, A. J., Sackett, P. R., Kuncel, N. R., Dahlke, J. A., & Beatty, A. S. (2022). Why women STEM majors are less likely than men to persist in completing a STEM degree: more than the individual. *Personality and Individual Differences*, *190*, 111532. https://doi.org/10.1016/j.paid.2022.111532

Koester, B. P., Grom, G., & McKay, T. A. (2016). Patterns of gendered performance difference in introductory stem courses. *Physics Education*, arXiv. https://doi.org/10.48550/ARXIV.1608.07565

Koller, M. (2016). robustlmm: An *R* package for robust estimation of linear mixed-effects models. *Journal of Statistical Software*, *75*(6). https://doi.org/10.18637/jss.v075.i06

Kuchynka, S. L., Eaton, A., & Rivera, L. M. (2022). Understanding and addressing gender-based inequities in STEM: research synthesis and recommendations for U.S. K-12 education. *Social Issues and Policy Review*, *16*(1), 252–288. https://doi.org/10.1111/sipr.12087

Lauer, S., Momsen, J., Offerdahl, E., Kryjevskaia, M., Christensen, W., & Montplaisir, L. (2013). Stereotyped: investigating gender in introductory science courses. *CBE—Life Sciences Education*, *12*(1), 30–38. https://doi.org/10.1187/cbe.12-08-0133

Lent, R. W., Sheu, H.-B., Miller, M. J., Cusick, M. E., Penn, L. T., & Truong, N. N. (2018). Predictors of science, technology, engineering, and mathematics choice options: A meta-analytic path analysis of the social–cognitive choice model by gender and race/ethnicity. *Journal of Counseling Psychology*, *65*(1), 17–35. https://doi.org/10.1037/cou0000243

Li, C. H., & Zafar, B. (2020). *Ask and you shall receive? Gender differences in regrades in college* (No. w26703; p. w26703). Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w26703

Lowe, P. A. (2019). Exploring cross-cultural and gender differences in test anxiety among U.S. and Canadian college students. *Journal of Psychoeducational Assessment*, *37*(1), 112–118. https://doi.org/10.1177/0734282917724904

Malespina, A., & Singh, C. (2022). Gender differences in grades versus grade penalties: Are grade anomalies more detrimental for female physics majors? *Physical Review Physics Education Research*, *15*(2), 020127. https://doi.org/10.1103/PhysRevPhysEducRes.18.020127

Malespina, A., & Singh, C. (2023). Gender gaps in grades versus grade penalties: Why grade anomalies may be more detrimental for women aspiring for careers in biological sciences. *International Journal of STEM Education*, *10*(1), 13. https://doi.org/10.1186/s40594-023-00399-7

Maloy, J., Kwapisz, M. B., & Hughes, B. E. (2022). Factors influencing retention of transgender and gender nonconforming students in undergraduate STEM majors. *CBE—Life Sciences Education*, *21*(1), ar13. https://doi.org/10.1187/cbe.21-05-0136

Maries, A., Whitcomb, K. M., & Singh, C. (2022). Gender inequities throughout STEM. *Journal of College Science Teaching*, *51*(3), 27–36.

Martin, D. C., & Arendale, D. A. (1992). Supplemental instruction: Improving first-year student success in high-risk courses. In *National Resource Center for the Freshman Year Experience* (2nd ed.). Columbia, SC: University of South Carolina.

Matz, R. L., Koester, B. P., Fiorini, S., Grom, G., Shepard, L., Stangor, C. G., ... & McKay, T. A. (2017). Patterns of gendered performance differences in large introductory courses at five research universities. *AERA Open*, *3*(4), 233285841774375. https://doi.org/10.1177/2332858417743754

McGee, E. O. (2020). Interrogating structural racism in STEM higher education. *Educational Researcher*, *49*(9), 633–644. https://doi.org/10.3102/0013189X20972718

Miller, L. D., Mitchell, C. E., & Van Ausdall, M. (1994). Evaluating achievement in mathematics: Exploring the gender biases of timed testing. *Education*, *114*(3), 436+.

Misra, R., & McKean, M. (2000). College students' academic stress and its relation to their anxiety, time management, and leisure satisfaction. *American Journal of Health Studies*, *16*(1), 41–51.

Miyake, A., Kost-Smith, L. E., Finkelstein, N. D., Pollock, S. J., Cohen, G. L., & Ito, T. A. (2010). Reducing the gender achievement gap in college science: a classroom study of values affirmation. *Science*, *330*(6008), 1234–1237. https://doi.org/10.1126/science.1195996

Mohammed, T. F., Nadile, E. M., Busch, C. A., Brister, D., Brownell, S. E., Claiborne, C. T., ... & Cooper, K. M. (2021). Aspects of large-enrollment online college science courses that exacerbate and alleviate student anxiety. *CBE—Life Sciences Education*, *20*(4), ar69. https://doi.org/10.1187/cbe.21-05-0132

Montolio, D., & Taberner, P. A. (2021). Gender differences under test pressure and their impact on academic performance: A quasi-experimental design. *Journal of Economic Behavior & Organization*, *191*, 1065–1090. https://doi.org/10.1016/j.jebo.2021.09.021

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, *109*(41), 16474–16479. https://doi.org/10.1073/pnas.1211286109

Nadile, E. M., Alfonso, E., Barreiros, B. M., Bevan-Thomas, W. D., Brownell, S. E., Chin, M. R., ... & Cooper, K. M. (2021). Call on me! Undergraduates' perceptions of voluntarily asking and answering questions in front of large-enrollment science classes. *PLOS ONE*, *16*(1), e0243731. https://doi.org/10.1371/journal.pone.0243731

National Science Foundation. (2019). *Women, minorities, and persons with disabilities in science and engineering* (NSF 19-304). Alexandria, VA: National Science Foundation, National Center for Science and Engineering Statistics. Retrieved January 22, 2022, from https://ncses.nsf.gov/pubs/nsf19304/

Niu, L. (2017). Family socioeconomic status and choice of STEM major in college: An analysis of a national sample. *College Student Journal*, *51*(2), 298–312.

Nord, C., Roey, S., Perkins, R., Lyons, M., Lemanski, N., Brown, J., & Schuknecht, J. (2011). *The Nation's report card: America's high school graduates (NCES 2011-462)*. Alexandria, VA: National Center for Education Statistics. Retrieved January 22, 2022. from https://nces.ed.gov/nationsreportcard/pdf/studies/2011462.pdf

Odom, S., Boso, H., Bowling, S., Brownell, S., Cotner, S., Creech, C., ... & C., J. (2021). Meta-analysis of gender performance gaps in undergraduate natural science courses. *CBE—Life Sciences Education*, *20*(3), ar40. https://doi.org/10.1187/cbe.20-11-0260

Osborne, J. W. (2001). Testing stereotype threat: does anxiety explain race and sex differences in achievement? *Contemporary Educational Psychology*, *26*(3), 291–310. https://doi.org/10.1006/ceps.2000.1052

Ost, B. (2010). The role of peers and grades in determining major persistence in the sciences. *Economics of Education Review*, *29*(6), 923–934. https://doi.org/10.1016/j.econedurev.2010.06.011

Pearson, M. I., Castle, S. D., Matz, R. L., Koester, B. P., & Byrd, W. C. (2022). Integrating critical approaches into quantitative stem equity work. *CBE—Life Sciences Education*, *21*(1), es1. https://doi.org/10.1187/cbe.21-06-0158

Perez-Felkner, L. (2018). Conceptualizing the field: higher education research on the STEM gender gap: conceptualizing gaps. *New Directions for Institutional Research*, *2018*(179), 11–26. https://doi.org/10.1002/ir.20273

Peterfreund, A. R., Rath, K. A., Xenos, S. P., & Bayliss, F. (2008). The impact of supplemental instruction on students in stem courses: results from San Francisco State University. *Journal of College Student Retention: Research, Theory & Practice*, *9*(4), 487–503. https://doi.org/10.2190/CS.9.4.e

Rabitoy, E. R., Hoffman, J. L., & Person, D. R. (2015). Supplemental instruction: the effect of demographic and academic preparation variables on community college student academic achievement in STEM-related fields. *Journal of Hispanic Higher Education*, *14*(3), 240–255. https://doi.org/10.1177/1538192714568808

Ramirez, G., & Beilock, S. L. (2011). Writing about testing worries boosts exam performance in the classroom. *Science*, *331*(6014), 211–213. https://doi.org/10.1126/science.1199427

Rask, K., & Tiefenthaler, J. (2008). The role of grade sensitivity in explaining the gender imbalance in undergraduate economics. *Economics of Education Review*, *27*(6), 676–687. https://doi.org/10.1016/j.econedurev.2007.09.010

Rath, K. A., Peterfreund, A. R., Xenos, S. P., Bayliss, F., & Carnal, N. (2007). Supplemental instruction in introductory biology I: enhancing the performance and retention of underrepresented minority students. *CBE—Life Sciences Education*, *6*(3), 203–216. https://doi.org/10.1187/cbe.06-10-0198

Rauschenberger, M. M., & Sweeder, R. D. (2010). Gender performance differences in biochemistry. *Biochemistry and Molecular Biology Education*, *38*(6), 380–384. https://doi.org/10.1002/bmb.20448

Robnett, R. D., Chemers, M. M., & Zurbriggen, E. L. (2015). Longitudinal associations among undergraduates' research experience, self-efficacy, and identity: Research Experience, Self-Efficacy, And Identity. *Journal of Research in Science Teaching*, *52*(6), 847–867. https://doi.org/10.1002/tea.21221

Rosenzweig, E. Q., Hecht, C. A., Priniski, S. J., Canning, E. A., Asher, M. W., Tibbetts, Y., ... & Harackiewicz, J. M. (2021). Inside the STEM pipeline: Changes in students' biomedical career plans across the college years. *Science Advances*, *7*(18), eabe0985. https://doi.org/10.1126/sciadv.abe0985

Ross, M. B., Glennon, B. M., Murciano-Goroff, R., Berkes, E. G., Weinberg, B. A., & Lane, J. I. (2022). Women are credited less in science than men. *Nature*, *608*(7921), 135–145. https://doi.org/10.1038/s41586-022-04966-w

Rosseel, Y. (2012). lavaan: An *R* package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Sabot, R., & Wakeman-Linn, J. (1991). Grade inflation and course choice. *Journal of Economic Perspectives*, *5*(1), 159–170. https://doi.org/10.1257/jep.5.1.159

Salehi, S., Cotner, S., Azarin, S. M., Carlson, E. E., Driessen, M., Ferry, V. E., ... & Ballen, C. J. (2019). Gender performance gaps across different assessment methods and the underlying mechanisms: the case of incoming preparation and test anxiety. *Frontiers in Education*, *4*, 107. https://doi.org/10.3389/feduc.2019.00107

Schmader, T. (2002). Gender identification moderates stereotype threat effects on women's math performance. *Journal of Experimental Social Psychology*, *38*(2), 194–201. https://doi.org/10.1006/jesp.2001.1500

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–646. https://doi.org/10.1214/aos/1176344136

Sekaquaptewa, D. (2019). Gender-based microaggressions in stem settings. *NCID Currents*, *1*(1), 1–10. https://doi.org/10.3998/currents.17387731.0001.101

Seymour, E., & Hewitt, N. (1997). *Talking about leaving: Why undergraduates leave the sciences.* Boulder, CO: Westview Press.

Shapiro, C. A., & Sax, L. J. (2011). Major selection and persistence for women in STEM. *New Directions for Institutional Research*, *2011*(152), 5–18. https://doi.org/10.1002/ir.404

Shapiro, R., Wisniewski, E., Kaeli, E., Cole, T., DiMilla, P., & Reisberg, R. (2016). Role of gender and use of supplemental instruction in a required freshman chemistry course by engineering students on their course grades and subsequent academic success. *2016 ASEE Annual Conference & Exposition Proceedings*, 26123. https://doi.org/10.18260/p.26123

Smit, R. (2012). Towards a clearer understanding of student disadvantage in higher education: Problematising deficit thinking. *Higher Education Research & Development*, *31*(3), 369–380. https://doi.org/10.1080/07294360.2011.634383

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, *35*(1), 4–28. https://doi.org/10.1006/jesp.1998.1373

Stanger-Hall, K. F. (2012). Multiple-choice exams: An obstacle for higher-level thinking in introductory science classes. *CBE—Life Sciences Education*, *11*(3), 294–306. https://doi.org/10.1187/cbe.11-11-0100

Steele, J., James, J. B., & Barnett, R. C. (2002). Learning in a man's world: examining the perceptions of undergraduate women in male-dominated academic areas. *Psychology of Women Quarterly*, *26*(1), 46–50. https://doi.org/10.1111/1471-6402.00042

Suárez, M. I., Dabney, A. R., Waxman, H. C., Scott, T. P., & Bentz, A. O. (2021). Exploring factors that predict STEM persistence at a large, public research university. *International Journal of Higher Education*, *10*(4), 161. https://doi.org/10.5430/ijhe.v10n4p161

Sue, D. W. (2010). *Microaggressions in everyday life: Race, gender, and sexual orientation*. Hoboken, NJ: Wiley.

Tarchinski, N. A., Rypkema, H., Finzell, T., Popov, Y. O., & McKay, T. A. (2022). Extended exam time has a minimal impact on disparities in student outcomes in introductory physics. *Frontiers in Education*, *7*, 831801. https://doi.org/10.3389/feduc.2022.831801

Theobald, E. (2018). Students are rarely independent: when, why, and how to use random effects in discipline-based education research. *CBE—Life Sciences Education*, *17*(3), rm2. https://doi.org/10.1187/cbe.17-12-0280

Theobald, E. J., Hill, M. J., Tran, E., Agrawal, S., Arroyo, E. N., Behling, S., ... & Freeman, S. (2020). Active learning narrows achievement gaps for underrepresented students in undergraduate science, technology, engineering, and math. *Proceedings of the National Academy of Sciences*, *117*(12), 6476–6483. https://doi.org/10.1073/pnas.1916903117

Thompson, S. K., Hebert, S., Berk, S., Brunelli, R., Creech, C., Drake, A. G., ... & Ballen, C. J. (2020). A call for data-driven networks to address equity in the context of undergraduate biology. *CBE—Life Sciences Education*, *19*(4), mr2. https://doi.org/10.1187/cbe.20-05-0085

Weaver, A. J., & Raptis, H. (2001). Gender differences in introductory atmospheric and oceanic science exams: multiple choice versus constructed response questions. *Journal of Science Education and Technology*, *10*(2), 115–126. https://doi.org/10.1023/A:1009412929239

Wei, T., & Simko, V. (2021). *R package "corrplot": Visualization of a Correlation Matrix* (0.92). Retrieved September 13, 2022, from https://cran.r-project.org/web/packages/corrplot/citation.html

Whitcomb, K. M., Cwik, S., & Singh, C. (2021). Not all disadvantages are equal: racial/ethnic minority students have largest disadvantage among demographic groups in both STEM and non-STEM GPA. *AERA Open*, *7*, 233285842110598. https://doi.org/10.1177/23328584211059823

Williams, M. M., & George-Jackson, C. E. (2014). Using and doing science: gender, self-efficacy, and science identity of undergraduate students in STEM. *Journal of Women and Minorities in Science and Engineering*, *20*(2), 99–126. https://doi.org/10.1615/JWomenMinorScienEng.2014004477

Wilton, M., Gonzalez-Niño, E., McPartlan, P., Terner, Z., Christoffersen, R. E., & Rothman, J. H. (2019). Improving academic performance, belonging, and retention through increasing structure of an introductory biology course. *CBE—Life Sciences Education*, *18*(4), ar53. https://doi.org/10.1187/cbe.18-08-0155

Witteveen, D., & Attewell, P. (2020). The STEM grading penalty: An alternative to the "leaky pipeline" hypothesis. *Science Education*, *104*(4), 714–735. https://doi.org/10.1002/sce.21580

Wright, C. D., Eddy, S. L., Wenderoth, M. P., Abshire, E., Blankenbiller, M., & Brownell, S. E. (2016). Cognitive difficulty and format of exams predicts gender and socioeconomic gaps in exam performance of students in introductory biology courses. *CBE—Life Sciences Education*, *15*(2), ar23. https://doi.org/10.1187/cbe.15-12-0246

Wu, Z., Spreckelsen, T. F., & Cohen, G. L. (2021). A meta-analysis of the effect of values affirmation on academic achievement. *Journal of Social Issues*, *77*(3), 702–750. https://doi.org/10.1111/josi.12415

Yeager, D. S., & Walton, G. M. (2011). Social-psychological interventions in education: they're not magic. *Review of Educational Research*, *81*(2), 267–301. https://doi.org/10.3102/0034654311405999

Zuur, A. F., Ieno, E. N., Walker, N., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. New York, NY: Springer. https://doi.org/10.1007/978-0-387-87458-6