Female In-Class Participation and Performance Increase with More Female Peers and/or a Female Instructor in Life Sciences Courses

E. G. Bailey,[†]* R. F. Greenall,[†] D. M. Baek,[‡§] C. Morris,[‖§] N. Nelson,[‡§] T. M. Quirante,[‡§] N. S. Rice,[‡§] S. Rose,[‡§] and K. R. Williams[¶§]

[†]Department of Biology, [‡]Department of Physiology and Developmental Biology, [‖]Department of Spanish and Portuguese, and [¶]Department of Microbiology and Molecular Biology, Brigham Young University, Provo, UT 84602

ABSTRACT

As we strive to make science education more inclusive, more research is needed to fully understand gender gaps in academic performance and in-class participation in the life sciences. Studies suggest that male voices dominate introductory biology courses, but no studies have been done on upper-level courses. Results on achievement gender gaps in biology vary and often conflict, and no studies have been done on the correlation between participation and academic performance gaps. We observed 34 life sciences courses at all levels at a large private university. Overall, males were more likely to participate than their female peers, but these gender gaps varied from class to class. Females participated more in classes in which the instructor called on most hands that were raised or in classes with more females in attendance. Performance gender gaps also varied by classroom, but female final course grades were as much as 0.2 SD higher in classes with a female instructor and/or a female student majority. Gender gaps in participation and final course grades were positively correlated, but this could be solely because female students are more likely to both participate more and earn higher grades in classes with many females in attendance.

INTRODUCTION

Women have lagged behind men in their representation in science, technology, engineering, and mathematics (STEM) for decades, but these gaps have been closing over time. The size of the gender[1] gap and rate at which it is closing varies by discipline, with the fewest females represented in engineering, physics, and computer science (West *et al.*, 2013; Board, 2018). The life sciences are often considered the most equitable of the STEM fields due to increasing representation of females, who are awarded 60% of life sciences bachelor's degrees and about half of doctoral degrees (West *et al.*, 2013; Board, 2018). Common explanations for these differences in female enrollment focus on the somewhat less quantitative and more "human-centric" nature of the life sciences compared with fields such as mathematics, computer sciences, and engineering (Ceci and Williams, 2010; Goulden *et al.*, 2011). A more recent study suggests that differences in representation can be explained by females experiencing a greater sense of belonging and higher self-efficacy in biology, chemistry, and mathematics compared with other STEM fields (Cheryan *et al.*, 2017).

[1]While biological sex is determined by physical/physiological characteristics, "gender" refers to socially constructed identities based on an individual's internal experience. In this study, we are primarily interested in gender differences, but we are limited to the information available to us as described in the *Methods* section. We will use "male" and "female" throughout the article for simplicity and clarity, but see footnotes throughout the paper for what these terms represent for each section of the study.

Sarah L. Eddy, *Monitoring Editor*

Submitted Dec 5, 2019; Revised May 20, 2020; Accepted May 21, 2020

CBE Life Sci Educ September 1, 2020 19:ar30

DOI:10.1187/cbe.19-12-0266

[§]Undergraduate student author.

*Address correspondence to: E. G. Bailey (liz_bailey@byu.edu).

Despite the initial appearance of gender equity within biological disciplines, researchers have pointed out the "leaky pipeline" phenomenon: female representation falls increasingly behind as careers progress (i.e., graduate school, postdoc positions, tenure-track appointments, authorship order, and invited talks; Wickware, 1997; Luckenbill-Edds, 2002; Schroeder *et al.*, 2013; West *et al.*, 2013; Sheltzer and Smith, 2014; Eddy and Brownell, 2016). Closing gender gaps in STEM fields is not only important for increasing opportunities for women to thrive and fill high-paying jobs, but researchers also point out that increased diversity benefits the entire field (Jehn *et al.*, 1999; Page, 2008). Shaw and Stanton (2012) identify two important bottlenecks during female academic careers that strongly influence this "leaky pipeline": choice of undergraduate major and application to faculty positions. In this study, we chose to focus on the female undergraduate experience, because female students' in-class experiences and course performances influence their likelihood of choosing and persisting in a major and field (Tinto, 1997; Rask and Tiefenthaler, 2008; Price, 2010; Rocca, 2010; Haldane *et al.*, 2012).

Verbal Participation in Class

As summarized in a meta-analysis (Jones and Dindia, 2004), the tendency of males to verbally dominate in classrooms is well documented from elementary through graduate education. However, many studies have found no in-class participation gender gaps, suggesting that other classroom or population characteristics likely influence the presence and size of gender gaps (Jones and Dindia, 2004; Rocca, 2010). For example, Canada and Pringle (1995) saw female participation decrease as the percentage of males in the classroom increased during the transition of a women's college into mixed-sex education, suggesting that the sex ratio in a classroom may be important. Results from studies investigating the effect of instructor gender on participation gender gaps vary widely, with some studies finding interactions and others finding no difference (Jones and Dindia, 2004). Fritschner (2000) found that females participated less than males in introductory courses but not in upper-level courses, suggesting course level may be important. Student characteristics other than gender could also have important effects on whether or not students participate. Jurik *et al.* (2013) found that high interest correlated with high participation, and the male students were more interested in the subject and thus participated more than the females.

Few studies on in-class participation gender gaps have been performed in the biological sciences specifically. Eddy *et al.* (2014) found that females participated less than their male peers in large introductory biology courses. A study in Norway found a similar participation gap in favor of males in introductory biology, even though this country has one of the highest ratings for gender equality in the world (Ballen *et al.*, 2017). However, both of these studies focused solely on introductory courses. Ballen *et al.* (2019) investigated the effect of a variety of classroom characteristics on gender disparities in participation and found that class size had the largest impact. Although not in a classroom, Carter *et al.* (2018) found that female scientists were less likely to ask questions during academic biology seminars. Thus, more studies are needed on in-class participation in the life sciences, especially in non-introductory undergraduate classes.

Performance

While gender achievement gaps are well documented in more quantitative STEM fields such as physics (Lorenzo *et al.*, 2006; Pollock *et al.*, 2007; Kost *et al.*, 2009; Kost-Smith *et al.*, 2010; Kreutzer and Boudreaux, 2012; Madsen *et al.*, 2013), fewer studies have been conducted regarding academic performance gaps in biology, and those that have been done give conflicting results. Ballen *et al.* (2018) found that females in lower-division biology classes underperformed compared with their male peers on high-stakes exams, but they earned higher scores on non–exam assignments. These gender gaps in favor of males on exam performance increased with class size. In a 13-year study in Michigan, female students averaged lower grades than their male peers in life sciences courses (Creech and Sweeder, 2012). Eddy *et al.* (2014) found a small yet consistent performance gap in favor of male students on exams in introductory biology courses. However, Lauer *et al.* (2013) found no gender gap in performance in biology classes when analyzing final grades and normalized learning gains on concept inventories. Wright *et al.* (2016) found that males performed significantly better than female peers as average Bloom's level increased on biology exams. Finally, Willoughby and Metz (2009) found mixed evidence for performance gender gaps in biology courses depending on how learning gains were calculated.

Researchers have also investigated whether instructor gender impacts the relative performance of male versus female students with conflicting results. In a business statistics class, students with the same gender as their instructor performed significantly better than students of the opposite gender (Haley *et al.*, 2007). Hoffmann and Oreopoulos (2009) looked at the interaction between student gender and instructor gender and its effect on course grades at the University of Toronto. They found a small effect of this gender interaction, but it was primarily due to male students underperforming in the social sciences when taught by a female instructor; performance gender gaps in math and science were negligible. A study conducted at the U.S. Air Force Academy found that females performed significantly worse than equally prepared male peers in math and science, but these gender gaps disappeared when there was a female professor (Carrell *et al.*, 2010). In that study, the highest-performing females seemed to benefit from female instruction the most, and males were not affected. In biology specifically, Eddy *et al.* (2014) found that achievement gaps in an introductory course were reduced in classes that were taught solely by female instructors.

Our Population

Gender gaps in STEM fields are likely to differ by population. Our study was conducted at a religious institution associated with the Church of Jesus Christ of Latter-Day Saints, so we were interested in how this culture would impact STEM gender gaps. A study conducted by Jensen and Jensen (1993) found that individuals with high religiosity were more likely to value the traditional female role in the home and that members of the Church of Jesus Christ of Latter-Day Saints were more likely to value this traditional role than Catholics or Protestants. Other studies have found that, in some populations, when female students attended a religious school or had a mother with more traditional gender ideologies, they were less likely to choose a male-dominated field such as STEM disciplines

(Rich and Golan, 1992; Steele and Barling, 1996). Similarly, Crawford (1978) found that females with more rigid views about gender roles were less likely to choose more male-dominated professions. Based on these studies, we hypothesized that gender gaps in STEM fields, including biology, might be larger in our population than in nonreligious institutions.

In this study, we aim to add to the growing body of literature about participation and performance gender gaps in undergraduate biology classrooms. Our study is unique, because we are quantifying both in-class participation and academic performance gender gaps, attempting to predict the size of these gender gaps, and including the participation gap as a possible predictor of the academic performance gap. We are also looking at all undergraduate class levels, rather than focusing on introductory courses, and focusing on a unique population with less diverse faculty and a student body that potentially holds more conservative ideas about gender roles.

Research Questions

1. Are male or female students more likely to participate in life sciences courses, and how large are in-class participation gender gaps?
2. What classroom characteristics predict the size of the in-class participation gap?
3. Are male or female students more likely to earn higher course grades in life sciences courses, and how large are these academic performance gender gaps?
4. Can classroom characteristics (including in-class participation gaps) be used to predict the size of the academic performance gender gap?
5. Are student gender, instructor gender, and an interaction between the two predictive of final course grades?

METHODS

Ethics Statement

Written consent was obtained from all course instructors, and permission for use of human subjects was obtained from the Brigham Young University Institutional Review Board (IRB).

Description of Classes

Data were collected from 34 classes (all taught by different instructors; see Supplemental Material for data set) in the College of Life Sciences of a large, private university: 10 during Winter semester 2015, eight in Spring 2015, six in Fall 2015, six in Winter 2016, and two in Spring 2016. Eighteen of the courses were taught by male instructors, and 16 were female-instructed. The gender ratios of our faculty sample (almost 50% female) do not represent the gender ratios of all life sciences faculty at the university (~15% female). However, we observed as many female instructors as possible, because we were specifically interested in the effects of instructor gender and wanted adequate statistical power. We included nine 100-level classes, five 200-level classes, sixteen 300-level classes, and four 400-level classes. Courses were only selected if they had a class size under 200, because it was difficult to accurately record participation events for classes larger than this. The largest class we included had about 160 attending students. Courses were observed at a variety of times throughout the semester based on student observers' schedules. Some were observed only early in the semester, some were observed only late in the semester, and some observations were spread out across the semester. This lack of consistent timing of observations is one weakness of our observation method.

Observation data were generally consistent across semesters, even though Spring terms are half as long as normal semesters. However, Winter semester of 2016 had larger class sizes than the rest of the semesters (one-way analysis of variance [ANOVA] with Tukey's posttest, $p = 0.004$), and Spring term of 2016 had the most active classrooms (class activity estimated as the number of times in a class period that the students talked to a neighbor or worked in groups; one-way ANOVA with Tukey's posttest, $p = 0.009$). It is unlikely this impacted our overall conclusions, but we included class size and classroom activity (as two variables: number of times the instructor asked the class a question and number of times the students worked in groups) as possible predictors in all regression analyses to account for these differences. Furthermore, we considered including semester as a random effect in all mixed model analyses, but it was ultimately not included, because it failed to improve the model.

Course syllabi were used to determine how much classroom participation was factored into the course grade for each class. Two researchers (E.G.B. and S.R.) read the participation policy for each course and coded it from one to six, with higher numbers indicating that the instructor placed a heavier emphasis on classroom participation:

1. Participation did not count toward the grade at all.
2. Participation was graded for a few important days during the semester.
3. Participation was graded periodically through the semester.
4. Participation points were collected daily through attendance, but these points were not formally considered in the course grade.
5. Participation points were collected daily, and these points were part of the course grade (participation did not have to be verbal: e.g., students were required to participate via writing or a response system).
6. Participation points were collected daily, and these points were part of the course grade (students needed to speak up in class to get the points).

If the two researchers classified a course differently, they discussed until they came to agreement. In the end, the degree to which participation was required did not show up as a significant predictor in any of our regression models (see *Results*), so our categorization scheme described here did not impact any of our conclusions. We also considered placing category 4 right after category 1 (because participation did not impact course grades). However, when the participation categories were ordered in that way, the degree to which participation was required was still not predictive in any of our models.

Participation

Data Collection. Most classes were observed three times, but the classes taught by instructors M15 and F11 were only observed twice due to scheduling difficulties. See the Supplemental Material for the data set. Pairs of student researchers (always one male and one female) sat in the back of the

classroom, each with a map of the room. The observers first labeled each occupied seat with the student's gender[2] (male or female), then throughout the class they marked every participation event for each student. Participation events included: hand raised, hand called on (student commented, asked a question, or answered a question), student called out without raising his or her hand ("call out," student commented, asked a question, or answered a question), student was called on by the instructor (nonrandom call), and student was called on by the instructor (random call). Students volunteering for activities that were not subject related were also recorded, but these were rare enough that we were not able to analyze these events further. As a limited way of quantifying the amount of student-centered pedagogy in the classroom, observers also recorded each time the instructor invited different kinds of participation: asking a question to the class or initiating some type of group work (think–pair–share or longer/larger group activities).

The two researchers recorded their observations separately during the class period. Immediately after the observation ended (so it was still fresh in their minds), the two researchers would meet and compare their classroom maps. For each difference in their map, the observers would discuss until they came to an agreement about what truly occurred. Then the data were officially recorded. Because observers modified their maps after coming to agreement, we cannot calculate interrater reliability statistics on the original data. Differences between observers were rare, but not having interrater reliability statistics is definitely a limitation of our study.

Gender Differences in In-Class Participation (Research Question 1). For statistical analyses in which classroom was the statistical unit ($n = 34$ classrooms), results from two or three observations were averaged to give one value for each class. For some analyses, all verbal participation events were pooled regardless of whether the student called out, was called on after raising his or her hand, or was selected by the instructor (termed "verbal participation"). A student raising his or her hand without getting called on was not considered verbal participation.

When classrooms were analyzed individually, the binomial exact test was used to test deviations from expected gender distributions. These expected distributions were determined by the average gender ratios of students in attendance for the two or three observations. Because the number of participation events was the statistical unit, participation results from the two or three observations were summed. Due to the large number of tests, we corrected p values for multiple comparisons (Benjamini and Hochberg, 1995) using a false discovery rate less than 0.05.

Predicting In-Class Participation Gender Gaps (Research Question 2). To compare male and female participation rates, we calculated the participation rate ratio (female/male) as the average rate of female participation divided by the average rate of male participation (i.e., [number of female participation events/number of female students]/[number of male participation events/number of male students]). Thus, a participation rate ratio = 1 suggests that male and female students were equally likely to participate in class (e.g., both males and females participated 1.2 times per class on average). A participation rate ratio <1 suggests that females were less likely to participate than males (e.g., if females participated 0.6 times per class on average, and males participated 1.2 times per class on average, the participation rate ratio = 0.5, suggesting females participated at half the rate of males). Finally, a participation rate ratio >1 suggests that females were more likely to participate than males (e.g., if females participated 1.2 times per class on average, and males only participated 0.6 times per class on average, the participation rate ratio = 2, suggesting females participated at twice the rate of males).

Linear mixed models were used with observation as the statistical unit. Models were selected using Akaike's information criterion corrected for small sample sizes (AICc). Models within an AICc of 2 were considered equivalent, and if they were within 2, the model with the fewest number of parameters was chosen as the best model. All analyses were performed using the linear mixed models function in IBM SPSS Statistics (v. 25). In brief, we generally used the method described by Theobald (2018) for all linear mixed models:

1. First, all possible fixed effects were included, and restricted maximum likelihood estimation was used to select random effects. The random effects included in the best-fitting model were retained for the rest of the analysis. (We considered including random intercepts for classroom and semester. In all cases, including a random intercept for each classroom improved the model, but including semester as a random effect did not. Thus, we only retained classroom as a random effect in our final models.)
2. Random effect inclusion (classroom) or exclusion (semester) was validated by calculating the intraclass correlation coefficient (ICC) of each random effect in an empty model.
3. Fixed effects were then selected using maximum likelihood estimation.
4. Finally, the best model was refit using restricted maximum likelihood estimation to get the most precise parameter estimates.

For all mixed models, only statistics from step 4 are included in the main text, but results of steps 1–3 are included in the Supplemental Material.

Academic Performance

Data Collection. For the same classes that were observed, we obtained final course grades from the university registrar's office. The data were de-identified except for gender[3] and ACT score. Three courses (M1, F1, and F16) were not included in the academic performance results, because they had fewer than

[2]Due to IRB limitations and the protection of students' privacy, classroom observers did not have access to self-reported gender identity or biological sex. Thus, student appearance was used to classify students as male- or female-presenting. This is obviously a limitation in our study, as we cannot be confident that our classification is accurate, nor can we account for the complexities of students' gender identities.

[3]In this instance, male and female classifications were obtained from the registrar's office and thus represent biological sex as included in official academic records. Again, this limits our study, because we cannot be confident that the biological sex on a student's academic record accurately or fully represents his or her self-reported gender.

five students of at least one gender enrolled in the course, and the registrar's office protects information for groups that small. Thus, $n = 31$ for analyses of performance.

Analysis of Academic Performance Gender Gaps in Individual Classes (Research Question 3). Average adjusted grades were obtained from estimated marginal means after a two-way analysis of covariance (classroom × student gender with ACT score as a covariate; $n = 1949$ students).

The independent-samples t test was then used to compare the adjusted final course grades of male versus female students for each class individually.

Predicting Classroom Gender Gaps in Academic Performance (Research Question 4). Performance gender gaps in adjusted grades were calculated for each classroom ($n = 31$): performance gender gap = average adjusted female grade – average adjusted male grade. When attempting to predict performance gender gaps, we nested our data by semester (unlike our other regression analyses, there is only a single value for each class in this case). Because grouping classes by semester did not explain variance in our data (semester ICC = 0.0 in an empty model), mixed model regression was not needed. Thus, stepwise multiple linear regression was performed to predict performance gender gaps. Variables were entered in the model if $p < 0.05$ (F-test), and variables were later removed if $p > 0.1$ (F-test).

Predicting Course Grades (Research Question 5). In this analysis, student was the statistical unit ($n = 1949$ students). Final course grades were not adjusted, because ACT was used as a predictor directly in the analysis. For our results to be more easily compared with other studies, grades on the four-point scale were standardized overall (based on the distribution of all classrooms combined) and grade z-scores were targeted using mixed model regression. We used the same linear mixed model protocol as described earlier for research question 2. Again, classroom was included as a random effect to account for grading style differences between classrooms.

RESULTS

Participation Gender Gaps (Research Question 1)

When we average results from all 34 observed life sciences classes, females were less likely to verbally participate in class than their male peers, with an average of 32% of males and 22% of females participating at least once (Figure 1A). Males were also more likely to be classified as "talkers" or verbally participate more than one time during a class period (Figure 1B). On average, males also had a higher verbal participation rate (number of verbal participation events/number of students; 0.77 average) compared with their female peers (0.43 average; Figure 1C). Our null hypothesis was that males and females are equally likely to participate; thus the fraction of female participation would be equal to the fraction of females in attendance. However, when we compared different types of participation with attendance fractions (Figure 2A), females were less likely to be heard verbally in the classroom, raise their hands, and call out than we predicted based on attendance. Because males were more likely to be talkers (see Figure 1B), we redid the analysis of Figure 2A to only count a participation event if it was the first time a student participated, thus ignoring the effect of talkers in the classroom. This was done separately for each participation type (i.e., students were only counted the first time they performed that specific participation type, but they may have previously participated in a different way). As shown in Figure 2B, we see that female students were still less likely to participate verbally overall, raise their hands, and call out compared with their male peers.

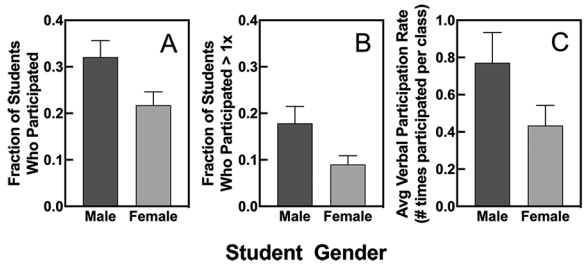

FIGURE 1. Males participate verbally more than females during class. (A) The average fractions of male vs. female students who verbally participated in class at least once during a class period (number of male students who participated/total number of males; number of female students who participated/total number of females) were compared by paired t test ($p < 0.0001$, $n = 34$ classes). (B) The average fractions of male vs. female students who verbally participated in class more than once during a period (number of male students who participated more than once/total number of males; number of female students who participated more than once/total number of females) were compared by paired t test ($p = 0.002$, $n = 34$ classes). (C) The average verbal participation rate was calculated as the number of verbal participation events divided by the number of students, and verbal participation of male vs. female students was compared by paired t test ($p = 0.0001$, $n = 34$ classes). Error bars represent SEM.

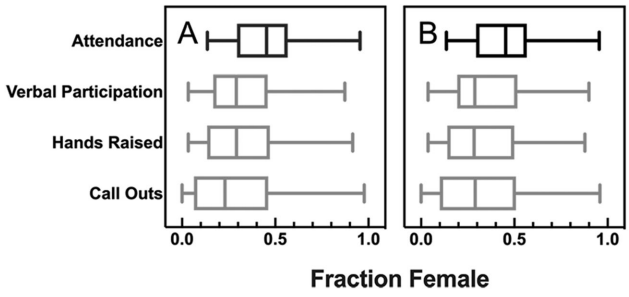

FIGURE 2. Females are less likely to participate verbally, raise their hands, and call out even when the effect of talkers is removed. (A) Average female fraction of students in attendance (dark) was compared with the average female fractions of verbal participation, hands raised, and call-outs (light) from classroom observations. The female fractions of all three participation types were distinguishable from the female fraction of students in attendance by repeated-measures one-way ANOVA with Dunnett's posttest ($p < 0.001$ and $n = 34$ classes for each comparison). (B) Data from A were recalculated to only count each student's first participation event. Female fractions of verbal participation, hands raised, and call-outs were each compared with the female fraction of students in attendance by repeated-measures one-way ANOVA with Dunnett's posttest ($p < 0.001$ and $n = 34$ classes for each comparison). Boxes indicate quartiles, and whiskers show full range of data.

We wondered whether differences in the likelihood of instructors to call on male versus female students could explain (at least partially) the gap in verbal participation. To determine whether instructors were treating male and female students equally, we calculated the call rate (average number of times called on/number of times hand raised) for male versus female students in each classroom. Overall, the average male call rate (89.7%) was indistinguishable from the average female call rate (89.5%) by paired *t* test ($p = 0.93$, $n = 34$ classrooms, unpublished data). Unequal treatment of males versus females could also be investigated by examining the rate at which instructors called on male and female students without the students raising their hands. True random call was only used in one of the classes we observed, but 14 out of the 34 instructors occasionally called on students nonrandomly without the students raising their hands. However, in all 14 classes, the number of nonrandom call events was too few to confidently assess whether the ratio of males and females called on deviated from expected frequencies based on attendance (unpublished data). Overall, we did not find evidence to suggest that our observed gaps in verbal participation could be attributed to explicit unequal treatment of males versus females by the instructors.

We found no interesting differences between males and females in terms of what types of verbal participation they offered. Both males and females were much more likely to ask and answer questions than they were to make a comment. Overall, 48.7% of female verbal participation was answering questions, 5.5% was commenting, and 45.8% was asking questions. This did not differ from the types of participation performed by male students: 46.5% answering, 5% commenting, and 48.5% asking (chi-square test for goodness of fit, $p = 0.36$).

When observed classes were analyzed individually, gender gaps in participation varied substantially from class to class. Female attendance is compared with female verbal participation for each class in Figure 3, with male-instructed classes shown in panel A and female-instructed classes shown in panel B. There were 11 classes in which females were less likely to participate than males (binomial exact test with adjustment for multiple comparisons, significance shown with asterisks in figure), one class in which females were more likely to participate than males (M4), and the null hypothesis was retained for the rest of the classes. Comparing female fraction of hands raised or female fraction of callouts to attendance yielded similar results (unpublished data), although differences between these types of participation and female attendance were less likely to be significant via the binomial exact test, because the overall number of events was lower than total verbal participation.

Predicting Participation Gender Gaps (Research Question 2)

Because participation gaps varied from class to class, we used linear mixed models to try and predict the gender gap in total

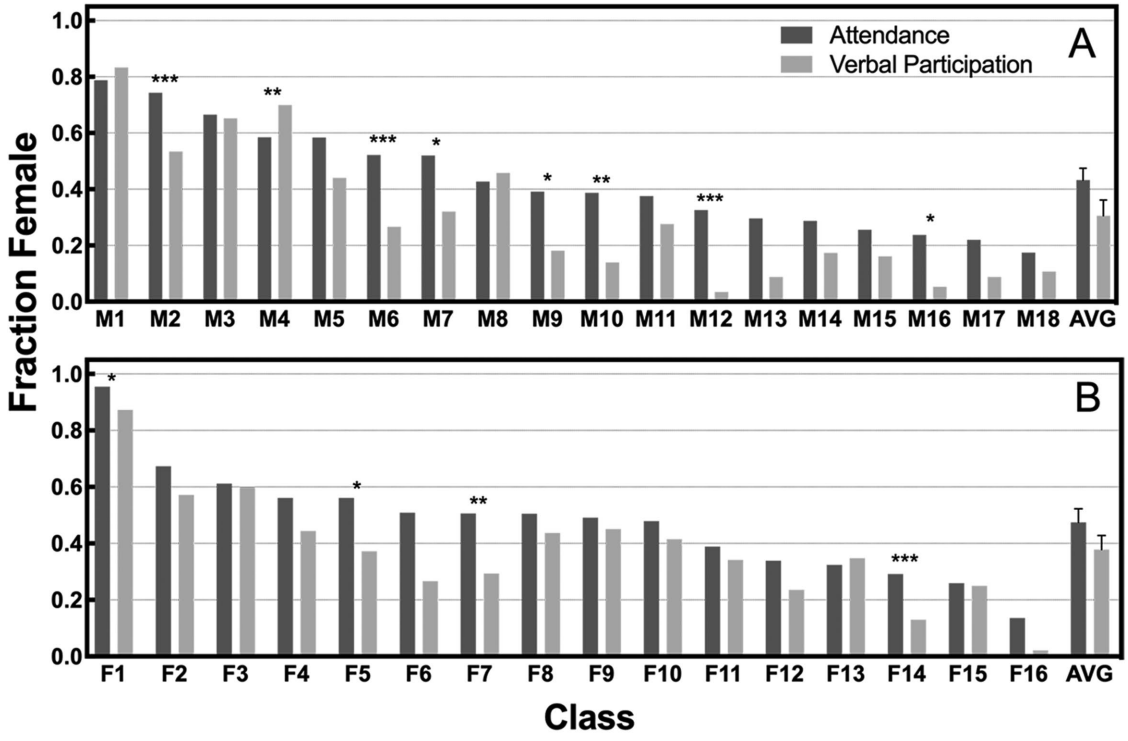

FIGURE 3. Participation gender gaps vary by classroom. For each classroom, the fraction of females in attendance was averaged across all observations, and the fraction of verbal participation events was calculated based on the sum of all observations (total number of female verbal participation events/total number of verbal participation events). Classes taught by a male are shown in A, and classes taught by a female are shown in B. The binomial exact test was performed for each classroom to determine whether gender ratios of verbal participation (light) deviated from expected frequencies (gender ratios in attendance; dark), and significant results are shown with asterisks after adjusting for multiple comparisons as described in the *Methods* (*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; ****$p < 0.0001$; $n = 15$–206 verbal participation events). Error bars on averages represent SEM.

TABLE 1. Student gender ratios predict verbal participation gender gaps, with more females predicting more female participation

Parameter	Relative variable importance	Included in best model?[a]	Regression coefficient ± SE
% Female students in attendance	0.74	Yes	0.716 ± 0.333
Upper-level course (300 or 400 level)	0.34	No	
Class size	0.30	No	
Female instructor	0.25	No	

[a]The best model also includes a random effect to allow for a random intercept for each class: (1|Classroom).

verbal participation. The gender gap in verbal participation was estimated for each observation by taking a ratio of the average female verbal participation rate over the average male verbal participation rate, with numbers lower than 1 suggesting the females participate less than males (see *Methods*). First, we considered course characteristics that instructors cannot control as possible predictors: class size, upper-level course (300 or 400 level), instructor gender, and the percentage of female students in attendance. As shown in Table 1, only the percentage of female students in attendance was retained as a fixed effect in the best model, with females being more likely to participate when there was a greater percentage of females in attendance (see *Methods* for description of statistical methods and Supplemental Tables S1 and S2 for detailed model selection results). A random intercept for each classroom was also included (validated by an ICC of 0.25). The resulting best model is shown in Equation 1.

$$\text{Verbal participation rate ratio}\left(\frac{\text{Fem}}{\text{Mal}}\right) = \text{Perc.Fem} + (1\,|\,\text{Classroom}) \quad (1)$$

Next, we considered course characteristics that are controlled by the instructor: overall call rate (average number of hands called on/number of hands raised), number of questions asked by the instructor, degree to which classroom participation is required for course credit (see *Methods*), and number of times students were asked to work in groups during class. As shown in Table 2, only overall call rate was included in the best model, and a high overall call rate predicted more female verbal participation (see *Methods* for a description of statistics and Supplemental Tables S3 and S4 for detailed model selection results). Again, a random intercept for classroom was included in the best model (see Equation 2).

$$\text{Verbal participation rate ratio}\left(\frac{\text{Fem}}{\text{Mal}}\right) = \text{Ov.Call.Rate} + (1\,|\,\text{Classroom}) \quad (2)$$

Thus, the percentage of females in attendance and the overall call rate by the instructor were the only course characteristics that could be used to predict verbal participation gender gaps. The relationship between these predictors and the verbal participation rate ratio (Fem/Mal) are shown visually in Figure 4 (predicted values from Equations 1 and 2 are plotted on the *y*-axis).

Academic Performance Gender Gaps (Research Question 3)

Final course grades were obtained for each observed classroom that had at least five males and five females enrolled, and results varied by classroom. As shown in Figure 5A, the difference in average male and female grades was only significant in two male-instructed classrooms when analyzed individually (M12 and M15, although it was close to significant in M17 as well). However, when data from all male-instructed classes were considered together (one-sample *t* test of Figure 5B), there was a small but consistent gender performance gap in favor of male students (the average female grade was 0.18 points lower than the average male grade on the four-point scale, which is 0.20 SD). This same effect was not seen in female-instructed classes (Figure 5, C and D), where no classrooms had a significant gender performance gap and no consistent trend was observable (the average female grade was 0.05 higher than the average male grade, which is 0.05 SD, but not significant by one-sample *t* test).

Predicting Academic Performance Gender Gaps (Research Questions 4 and 5)

Due to the variability from classroom to classroom, we attempted to predict classroom performance gaps using multiple linear regression. Our target variable was the classroom performance gender gap, calculated as female average grade minus the male average grade after adjusting grades for ACT score (see *Methods*); thus, a positive number indicates females outperform male peers, while a negative number indicates females underperform compared with their male peers. Possible predictors included course characteristics instructors cannot control (upper-level course, class size, percentage of

TABLE 2. Increased overall call rate positively predicts female verbal participation

Parameter	Relative variable importance	Included in best model?[a]	Regression coefficient ± SE
Overall call rate (number of hands called on/number of hands raised)	0.93	Yes	0.995 ± 0.375
Number of times students worked in groups	0.46	No	
Number of questions asked by instructor	0.36	No	
Participation required for course credit	0.27	No	

[a]The best model also includes a random effect to allow for a random intercept for each class: (1|Classroom).

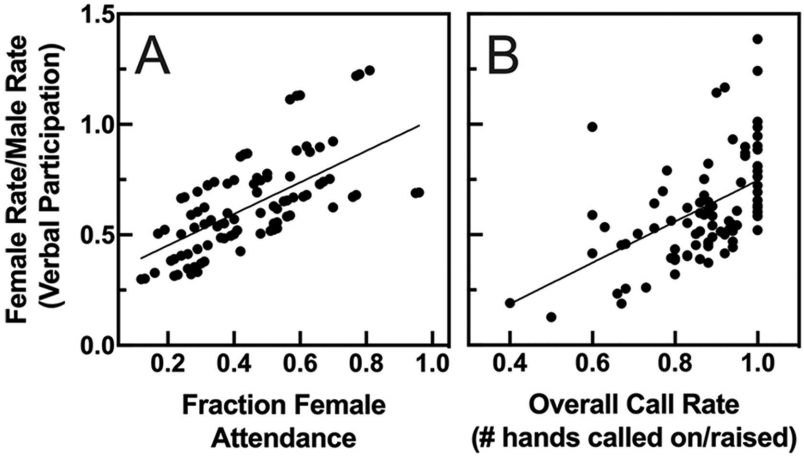

FIGURE 4. Higher female attendance and overall call rate predict greater female verbal participation. The verbal participation rate ratio was calculated as the female participation rate (number of verbal events/number of students) over the male participation rate. This rate ratio was predicted separately using classroom characteristics instructors have no control over (A) or variables instructors do control (B). The best linear mixed models included classroom as a random effect as well as fraction of females in attendance (A) or overall call rate (average number of times called on/number of times hand raised; B) as fixed effects. These two models generated the predicted verbal participation rate ratio (Fem/Mal; *y*-axis). The regression coefficient of fraction females in attendance was 0.716 ± 0.333, and the coefficient for overall call rate was 0.995 ± 0.375 (mean ± SE, $n = 100$ observations from 34 classrooms).

females in attendance, and instructor gender), course characteristics under the control of the instructor (overall call rate, number of instructor questions, degree to which participation is required in the course grade, and number of times students worked in groups), and verbal participation gaps. A stepwise regression method was used, allowing predictors to be added to the model when $p < 0.05$ and removed from the model when $p > 0.1$. The first predictor to significantly predict academic performance gaps was having a female instructor (see Table 3, model 1). This model explained ~36% of the variation in classroom performance gaps, with a female instructor positively predicting female performance. When the percentage of females in attendance was added as a predictor, the model could now explain ~57% of the variation in performance gaps (see Table 3, model 2; $n = 30$), and more females in attendance predicted better female performance. These results are represented visually in Figure 6, with male-instructed classes shown as dark circles and female-instructed classes shown as light diamonds. No other predictors were able to significantly predict performance gender gaps, leaving Equation 3 as our final model.

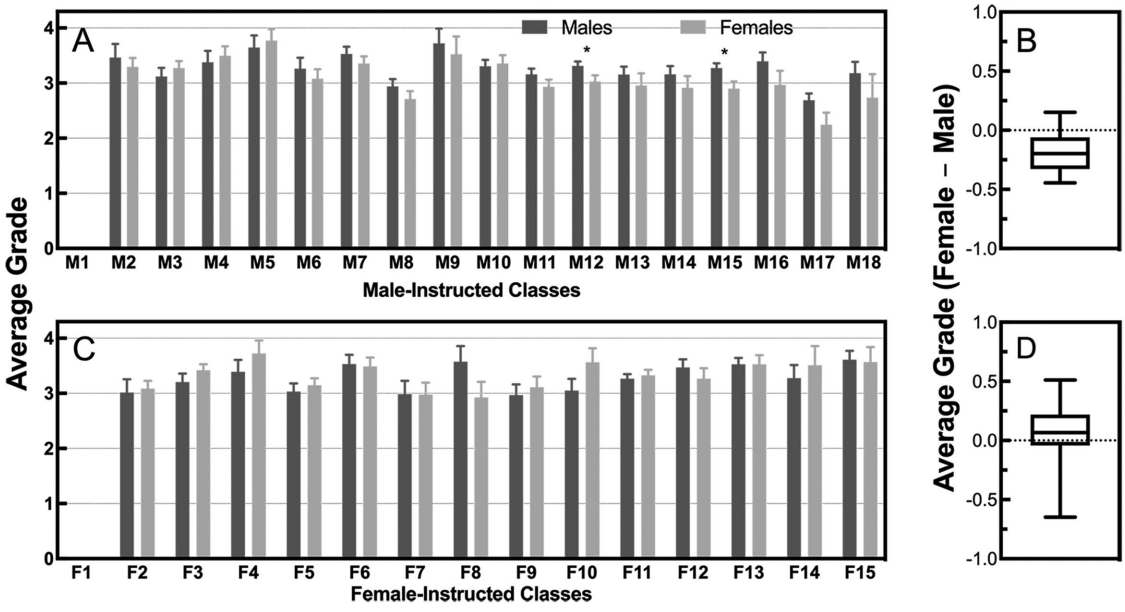

FIGURE 5. On average, male-instructed classes exhibit a performance gender gap in favor of males, but female-instructed classes do not. (A) Average grades (adjusted for ACT score) are shown for male (dark) and female students (light) in each observed male-instructed class (instructor ID numbers on *x*-axis). When analyzed individually, two classes had significant gender gaps (independent-samples *t* tests; M12: $p = 0.04$, $n = 182$; M15: $p = 0.02$, $n = 155$; rest of classes: $n = 19–97$ students). (B) Performance gender gaps in all male-instructed classes were averaged, and females significantly underperformed their male peers (one-sample *t* test, $p = 0.001$, $n = 17$ classes). (C) Average ACT-adjusted grades are shown for the female-instructed classes, and no performance gender gaps were significant when classes were analyzed individually (independent-samples *t* tests, $n = 21–190$ students). (D) When all female-instructed classes are averaged, there is no significant achievement gap (one-sample *t* test, $p = 0.48$, $n = 14$ classes). Error bars represent SEM.

TABLE 3. Results of stepwise multiple linear regression with performance gap as target

Model	R^2	Adjusted R^2	Significance (change in R^2)	Variable	B (coefficient)	SE_B	β (standardized coefficient)	p value
1	0.380	0.358	<0.0001	(Intercept)	0.106	0.053	0.617	0.054
				Female instructor	0.291	0.070		<0.0001
2	0.602	0.573	0.001	(Intercept)	–0.240	0.099	0.538	0.022
				Female instructor	0.254	0.058	0.478	<0.0001
				Percent female in attendance	0.750	0.193		0.001

$$\text{Gender gap in academic performance} = \text{Fem.Inst} + \text{Perc.Fem} \quad (3)$$

The regression results of Table 3 were obtained after excluding instructor F8. The performance gender gap in this instructor's classroom was 2.5 SDs from the mean for all instructors, so it could be considered an outlier under some definitions. When instructor F8 is included in the multiple linear regression analysis, female instructor and percentage of females in attendance were still the only two variables that significantly predicted gender gaps in performance, but the models explained less of the variance (see Supplemental Table S5 for adjusted R^2 values and other detailed results when instructor F8 is included; $n = 31$ classrooms). Thus, we feel confident that instructor gender and percentage of females in attendance are significant predictors of performance gender gaps in our population.

There are three possible explanations for the closing of a performance gender gap in courses with a female instructor and/or a large percentage of females in attendance seen in Figure 6: (1) female performance benefits, (2) male performance suffers, or (3) both are true. We took two different approaches to distinguish among these three possibilities. First, we repeated the multiple linear regression just described, but we predicted ACT-adjusted male grades and ACT-adjusted female grades separately (as opposed to predicting the difference between the two). Female instructor and percentage of females in attendance were both significant predictors of increased female performance ($p = 0.045$ and 0.034, respectively, adjusted $R^2 = 0.242$), but a model with these two predictors could not significantly predict male performance (see full results in Supplemental Table S6).

Second, we performed linear mixed models to predict student course grades (converted to z-scores). Possible predictors included ACT score (ACT), instructor gender (Fem.Inst), student gender (Fem.Stud), and an interaction between instructor gender and student gender (Fem.Inst*Fem.Stud). As shown in Table 4, ACT score and an interaction between instructor and student gender were retained in the best model (see *Methods* for description of statistics and Supplemental Tables S7 and S8 for detailed model selection results). A random intercept for classroom was also included to account for student grouping within classes (even though the ICC was only 0.05, including this random effect still improved our model significantly). The resulting best model is shown in Equation 4.

$$\text{Student course grade } z\text{-score} = \text{ACT} + \text{Fem.Inst*Fem.Stud} + (1 \mid \text{Classroom}) \quad (4)$$

Predicted course z-scores from this model are shown visually in Figure 7. Based on this analysis, female students benefit by having a female instructor, but male students are not affected. Based on the results of Table 4 and Figure 7, we conclude that the reduction in the performance gender gap with a female instructor and/or more females in attendance is due to the female students benefiting rather than any impairment of male students.

DISCUSSION

Participation (Research Questions 1 and 2)

We found that females participated less than their male peers on average in life science classes at a large private university (Figures 1 and 2), but the results did vary by classroom (Figure 3). Our results generally matched those found by Eddy *et al.* (2014) and Ballen *et al.* (2017), in that females participated less than expected on average, and some classes had larger gaps than others; however, we saw a smaller average gap in verbal participation (~10% less than the percentage of females attending; see Figures 2 and 3) compared with these two studies (~20% less than percentage of females attending). We had hypothesized that we would see larger participation gaps due to our religiously conservative population, so this surprised us.

There are at least two possible explanations as to why our gender gap effect size was smaller than previous studies. First, it is possible that gender gaps in participation are simply smaller

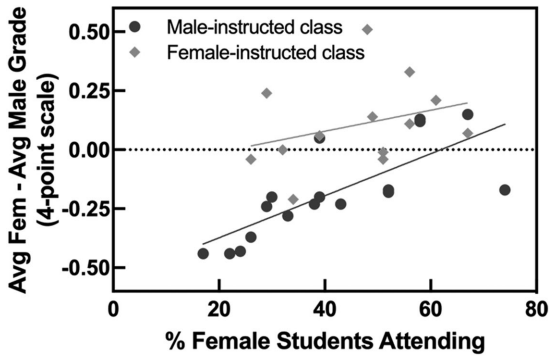

FIGURE 6. Performance gender gaps are predicted by instructor gender and percentage of females in attendance. The gender gap in performance was calculated for each class as female average grade minus the male average grade (ACT adjusted). These performance gaps are plotted against percent female students in attendance, with male-instructed classes shown with dark circles and female-instructed classes shown with light diamonds. Simple linear regression was performed on male- and female-instructed classes separately to yield the dark and light lines. The slopes of these best-fit lines were indistinguishable ($p = 0.30$), but the intercepts were significantly different ($p = 0.0002$, $n = 30$). If instructor F8 (an outlier) had been included, the slopes would still be equivalent ($p = 0.28$), and the female-instructed classes would still have a higher intercept ($p = 0.015$, $n = 31$).

TABLE 4. Student performance (course grade *z*-scores) can be predicted by ACT score and an interaction between student gender and instructor gender

Parameter	Relative variable importance	Included in best model?[a]	Regression coefficient ± SE
ACT score	1.00	Yes	0.076 ± 0.006
Instructor gender*student gender	0.95	Yes	
*(Reference: Female instructor*female student)*			
Male instructor*male student			–0.062 ± 0.107
Male instructor*female student			–0.253 ± 0.110
Female instructor*male student			–0.054 ± 0.069
Student gender	0.51	No	
Instructor gender	0.50	No	

[a]The best model also includes a random effect to allow for a random intercept for each class: (1|Classroom).

in our population. This would contradict our original hypothesis, which was informed by studies that suggest that religious education and deeply held beliefs about traditional gender roles can keep females out of STEM fields (Rich and Golan, 1992; Steele and Barling, 1996). Perhaps the focus the Church of Jesus Christ of Latter-Day Saints places on public speaking in church services (for men, women, and children) makes females more comfortable speaking up in class than is common in other cultures (Ludlow, 1992). Qualitative research investigating female students' experiences would be needed to investigate this possibility. Second, our smaller gap could simply be attributable to differences in our methodology compared with previous studies. We observed smaller classrooms (fewer than 160 attending students) than previous studies (up to 900 students), and we observed all course levels, not just introductory classes. While class size was not a significant predictor in our models of verbal participation gaps, it might have been if we had observed large enough classes. However, "upper level" was not a significant predictor of participation gender gaps, so we did not find evidence to support the idea that introductory courses have larger gender gaps in participation.

Previous studies on participation gender gaps in biology did not account for talkers, or students who participate multiple times in a class period. If males are more likely to be talkers than females (as we found in Figure 1B), then it is possible that participation gender gaps found previously were caused by only a few students. However, as shown in Figure 2B, our observed gender gaps in participation were still significant (and of comparable size) when students were only counted the first time they participated. Thus, the tendency of female voices to be heard less than voices of their male peers in a biology classroom cannot be accounted for by a few vocal students only.

As our study was only observational, it is difficult to provide commentary on causal factors behind these gender gaps in participation. However, we tried measuring different variables related to classroom environment and teaching methods to see whether they were related to participation gender gaps. First, we found no evidence that instructors were explicitly favoring male students, as on average, instructors called on ~90% of students regardless of student gender.

Second, we looked at class characteristics that instructors have no control over (course level, class size, percentage of females in attendance, and the instructor's gender), and the percentage of females in attendance was a positive predictor of female participation (Table 1 and Supplemental Table S2). This complements the findings of a study that saw female participation go down after transitioning from an all-girls school to mixed-sex education (Canada and Pringle, 1995), but to our knowledge, this is the first time this effect has been shown in biology courses specifically. Perhaps we saw this when other studies did not simply because we had more variance in our attendance ratios than past studies (Eddy *et al.*, 2014; Ballen *et al.*, 2017, 2019). It is also possible that having female peers is more important for female students in more conservative populations. Course level, class size, and instructor gender were not significant predictors of participation gaps in our population. Eddy *et al.* (2014) and Ballen *et al.* (2019) similarly found no effect of instructor gender on participation gender gaps. Interestingly, we did see significantly more participation overall in female-instructed classrooms (unpublished data), but this had no impact on the participation gender gap, because male and female students were both more likely to participate with a female instructor. Ballen *et al.* (2019) reported that class size had the largest impact on participation gender disparities (with females participating less in large classrooms). They had greater variability in class size in their sample than we did, so

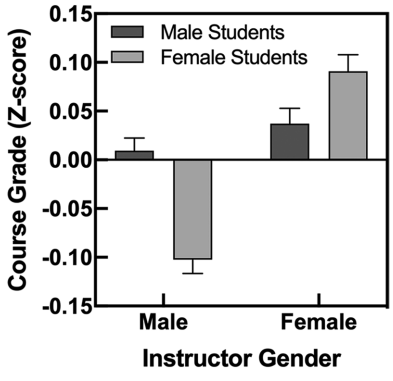

FIGURE 7. Female instruction is predictive of higher course grades for female students. Course letter grades were obtained for 1949 students in 31 classes, converted to the four-point scale, then converted to standardized scores (*z*-scores) based on the distribution of all grades overall. These *z*-scores were then modeled using a linear mixed model containing ACT score and an instructor gender*student gender interaction as fixed effects and classroom as a random effect. This model was then used to generate predicted course grade *z*-scores (*y*-axis), presented here by instructor and student gender. Error bars represent SEM.

we might have found the same thing if we had sampled larger classrooms.

Third, we assessed classroom variables that instructors do control (overall call rate, degree to which participation is required, number of questions instructors ask, and number of times students worked in groups). Only overall call rate was a significant predictor of gender gaps in participation: Increased female participation was predicted by the instructor calling on more of the raised hands (Figure 4, Table 2, and Supplemental Table S4). This could suggest that females are more likely to raise their hand and speak up if they do not have to compete to be called on, that is, if there is a high probability the instructor will choose them. We did not see significant reductions in participation gender gaps if classrooms were more student centered; however, we counted the number of times students were asked questions and/or told to work in groups rather than quantifying the amount of time students actually spent talking and working on such tasks. It is possible that we might have obtained different results had we evaluated teaching methods more thoroughly, and this should be done in future studies. Ballen *et al.* (2019) accounted for teaching methods by quantifying both the number of student–instructor interactions per class and the diversity of those interactions, but they likewise did not find significant impacts on participation gaps.

Other elements of classroom culture that are difficult to measure and quantify could also contribute to differences in the ratio of male and female voices heard in biology classrooms. For example, we cannot rule out effects of implicit gender biases of instructors and students. Implicit bias theory suggests that individuals do not always have conscious, intentional control over what motivates their actions (Greenwald and Krieger, 2006). For example, in a randomized double-blind experiment, science faculty rated a male job applicant as more competent than a female applicant, even though the applications were completely identical other than gender (Moss-Racusin *et al.*, 2012). In that study, pre-existing subtle bias against women was predictive of less support for the female candidate. In our study, it is possible that instructors' implicit gender biases contributed to a classroom culture that decreased female students' sense of belonging, or female students' own implicit gender biases could have made them feel like they did not belong. Sense of belonging has been found to be associated with both the degree to which instructors encourage participation and the degree to which students participate in class (Finn and Cox, 1992; Goodenow and Grady, 1993; Freeman *et al.*, 2007). We were not able to assess our participants' implicit gender biases or their sense of belonging; however, these subtle factors should be investigated in future studies on gender gaps in in-class participation.

Academic Performance (Research Questions 3–5)

Gender differences in final course grades were not as dramatic as those in participation in our population, and they varied by classroom (Figures 5 and 6). Thus, we were interested in class characteristics that could predict the size of gender gaps. Variables that reflect instructors' teaching style (overall call rate, the number of times instructors posed questions to the class, the number of times instructors had students work in groups, and participation grading policies) were considered as possible predictors of classroom performance gender gaps, but none significantly improved the regression model (Table 3). Other choices instructors make about their course could impact gender differences in achievement. For example, Wright *et al.* (2016) found that males perform better than females on exams that include items requiring higher-order cognitive skills, whereas they found no gender gap on exams merely testing lower-order cognitive skills. The weighting of high-stakes assignments (e.g., exams) versus low-stakes assignments (e.g., homework) has also been found to affect achievement by gender, with males earning more high-stakes points and females outperforming on low-stakes points (Ballen *et al.*, 2018). In this study, we were limited to collecting only final course grades from the registrar's office. In addition, we did not have information about what types of exams were given in each class, nor did we know the breakdown of the weighting of different types of assignments for final grade calculations. It is possible that these variables might have been able to predict performance gender gaps had we been able to include them in our regression analysis, and we would like to test this more explicitly in the future.

We did find that gender gaps in performance were influenced by instructor gender (Figures 5–7 and Tables 3 and 4). In male-instructed classes, female students earned final course grades 0.1–0.2 SD lower than their male peers on average; the size of the gap varied based on whether gender gaps were calculated by classroom (Table 3 and Figures 5 and 6; 0.2 SD) or student grades were targeted with ACT, gender, and classroom as regression factors (Table 4 and Figure 7; 0.1 SD). Lauer *et al.* (2013) found no gender gap in achievement in final introductory biology course grades, but the class they used was taught by a female instructor, so our results align well. Our effect size is comparable to existing studies that did see gender gaps in performance. Carrell *et al.* (2010) found that female students scored 0.15 SD lower than males when randomly assigned to a math or science class taught by a male instructor. Eddy *et al.* (2014) reported a 0.2 SD gap in favor of males on introductory biology exams.

Here we report that having a female instructor closed the performance gender gap, with females performing equally with (or perhaps even outperforming) male peers on final course grades (Figures 5–7). Other studies have also seen female student scores increase with female instruction, but with effect sizes smaller than ours: 0.05–0.1 SD (Carrell *et al.*, 2010; Eddy *et al.*, 2014). Carrell *et al.* (2010) found that having a female instructor closed achievement gender gaps, because female instruction had a negative impact on male students' performance. Conversely, in our population, closing of the performance gender gap was attributed to female students performing better (~0.2 SD) with a female instructor compared with a male instructor, rather than male students being negatively affected (Figure 7, Table 4, and Supplemental Table S6). In fact, if anything, male performance also increased in female-instructed courses.

There are at least three explanations for the observation that females earned higher final course grades when taught by a female instructor. First, as discussed earlier, there is evidence that males are favored on higher-level cognitive skills and high-stakes assignments (Wright *et al.*, 2016; Ballen *et al.*, 2018). The instructor gender effect we observed could be explained if female instructors' exams were aimed at lower-level cognitive skills and/or less heavily weighted in final grade calculations than male instructors' exams. While we think this is unlikely

and find no support for this idea in the literature, we cannot exclude it as a possibility, because we do not have data about the exams given in each course (their cognitive difficulty or their weight in final course grades).

Second, our instructor gender effect could be explained if male professors were grading female students more harshly than male students. Implicit bias theory suggests that this could be done subconsciously, as with instructors rating a male applicant more capable than a female applicant (Moss-Racusin *et al.*, 2012). However, in that study, even the female professors exhibited implicit gender bias in favor of male students, suggesting that unfair grading could also be found in female-instructed classrooms. It is possible that the male professors in our population have more implicit bias against female students than the female instructors do. As discussed earlier, more traditional gender roles are likely more valued in our population compared with others, and it is possible that males hold these values more than females. Furthermore, we do not know what proportion of work was subjectively graded, whether teaching assistants helped with grading, whether grading was blind (without knowing student names), or the implicit gender bias tendencies of each instructor. In the future, this type of data would be necessary to determine whether increased female performance in female-instructed courses could be attributed to differences in grading procedures.

Third, perhaps females perform better with female instructors because the instructor serves as a positive role model for the female students, increasing belonging and mitigating stereotype threat. Research suggests that feeling a sense of belonging in a group is related to increased student motivation, self-efficacy, task utility, and academic engagement (Freeman *et al.*, 2007; Walton *et al.*, 2012; Brownell *et al.*, 2014; Wilson *et al.*, 2015; Lewis *et al.*, 2016); that females feel less belonging than males in at least some STEM fields (Lewis *et al.*, 2017); and that females can feel less belonging when they do not see women represented (Murphy *et al.*, 2007). In a lab setting, individuals underperformed if they were the only member of the group of a particular low-status race or gender, suggesting that a lack of belonging can also impact performance (Sekaquaptewa and Thompson, 2002). If students seeing (or not seeing) others "like them" can impact their sense of belonging, then female students would likely benefit from having female instructors to act as role models. This might be especially important in a population like ours, in which female STEM professors are in the vast minority and women are less likely to work outside the home (Leamaster and Subramaniam, 2016).

Dasgupta (2011) proposed the "stereotype inoculation model," suggesting that in-group (in this case, female) experts can act like "social vaccines" to increase belonging in their group members (female students) and to "inoculate" against stereotype threat. Stereotype threat is the fear of being reduced to a bad stereotype that exists about a group (Steele, 1997), such as women being less capable in STEM. There is evidence that stereotype threat can obstruct female achievement in STEM, that female performance decreases if they are reminded of stereotypes, and that female performance can increase when stereotype threat is addressed and mitigated using values-affirmation tasks (Spencer *et al.*, 1999; Miyake *et al.*, 2010; Shapiro and Williams, 2012). However, a recent study done in introductory science classes found little evidence of stereotype threat endorsement or a benefit to values-affirmation writing tasks, but neither did they find a gender gap in achievement (Lauer *et al.*, 2013). Thus, it remains possible that stereotype threat might still be an issue in our population, for whom we do see achievement gaps in some courses. While female instructors could reduce stereotype threat for female students (Young *et al.*, 2013), there is also evidence that the degree to which female instructors embody the stereotype impacts their ability to do so (Cheryan *et al.*, 2011). This could explain why there is still variability from class to class. Because female professors make up only about 15% of the full-time life sciences faculty at this institution, we hypothesize that having a female instructor is important for the female students and likely the reason female instruction is such a strong predictor of female performance in our study population. Future research should explicitly focus on female students' sense of belonging and how it is affected by having female instructors.

As with our participation gender gaps, gender gaps in final course grades were reduced in classrooms with more females in attendance (Figure 6). Having a lot of female peers likely increases female student performance for many of the same reasons as having a female instructor does so, by increasing female students' sense of belonging and reducing stereotype threat. Peers may be especially helpful in a population like ours, where female instructors can be rare. Even if the instructor at the front of the room is male, females can potentially see each other as role models and as a sign that others in the field look "like themselves," that they "fit in."

To our knowledge, no previous studies have investigated a potential correlation between participation gender gaps and performance gender gaps. We wondered whether classes in which female voices were heard less would also be the classes in which females earned lower grades. The participation gender gap (participation rate ratio [Fem/Mal]) was positively correlated with the gap in final course grades ($r = 0.355$, $p = 0.049$, $n = 31$), yet the participation rate ratio was not a significant predictor of performance gaps in the linear mixed models (Table 3). Interestingly, having more females present in the classroom was a predictor of both increased female participation and performance (Tables 1 and 3), so that suggests that we see increased female performance in the same classes in which females feel more comfortable speaking up. Perhaps the attendance ratio explained so much variance in academic performance gaps that participation gaps had no additional statistical explanatory power. Nevertheless, this result still supports the hypothesis that females perform better in classes in which they feel a strong sense of belonging and that feelings of belonging can be strengthened by having many female peers. Future studies will need to be done to determine whether interventions to increase female participation (especially in classrooms in which females make up the minority) also increase female performance.

Summary

We studied gender gaps in participation and performance in undergraduate life sciences classes at all levels at a large, private university. In our population, we found that female voices are less likely to be heard in most classes and that these participation gender gaps remain significant even when the effect of talkers is eliminated (Figures 1 and 2). These in-class

participation gender gaps vary in size by classroom and are smaller in classrooms with more female students and in which students are more likely to get called on when they raise their hands (Figures 3 and 4 and Table 2). We evaluated achievement by collecting final course grades by gender, and we used ACT scores to consider college preparation. On average, male-instructed courses exhibit a significant gender gap in performance in favor of male students, while the average gap in female-instructed courses is not distinguishable from zero (Figure 5). In regression analyses, having a female instructor and many female peers are positive predictors of female performance (Figures 6 and 7 and Tables 3 and 4). Participation gender gaps are correlated with performance gender gaps, but this could be explained entirely by both being impacted by the student gender ratio. We propose that female instructors and peers can increase female students' sense of belonging and mitigate stereotype threat, but this hypothesis will need to be explicitly tested in future studies.

Implications

As we strive for greater gender equity in biology education, our results suggest that the following considerations may be important to increase female success. First, we provide evidence that female professors have the potential to increase female student performance. In institutions like ours, where female faculty are very much in the minority, efforts could be made to help faculty demographics more closely match student demographics. Male professors might also consider hiring female teaching assistants or inviting female guest speakers to give female students more role models to whom they can relate. Second, this study suggests that female students earn lower grades and are less likely to participate when they are in the minority. Obviously, instructors cannot control the gender ratios in their classroom, but they can be mindful of gender ratios in any groups they form in the classroom. There is a lot of literature about gender composition in small groups, some of it contradictory. Results from many studies support the idea that female performance and participation suffer when females are in the minority (Lee, 1993; Myaskovsky *et al.*, 2005; Harskamp *et al.*, 2008), while other studies suggest that females may prefer the minority and/or not do well in small groups composed only of females (Gnesdilow *et al.*, 2013; Zhan *et al.*, 2015). More research is warranted on this topic, especially in the life sciences. Finally, females were more likely to participate in our study when the instructor called on a greater percentage of hands raised. This suggests that instructors should strive to create a classroom environment that values all voices rather than being competitive.

REFERENCES

Ballen, C. J., Aguillon, S. M., Awwad, A., Bjune, A. E., Challou, D., Drake, A. G., … & Goldberg, E. E. (2019). Smaller classes promote equitable student participation in STEM. *BioScience*, *69*(8), 669–680.

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.

Ballen, C. J., Danielsen, M., Jørgensen, C., Grytnes, J.-A., & Cotner, S. (2017). Norway's gender gap: Classroom participation in undergraduate introductory science. *Nordic Journal of STEM Education*, *1*(1), 262–270.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300.

Board, N. S. (2018). *Science and engineering indicators 2018*. Alexandria, VA: National Science Foundation.

Brownell, S. E., Freeman, S., Wenderoth, M. P., & Crowe, A. J. (2014). BioCore Guide: A tool for interpreting the core concepts of *Vision and Change* for biology majors. *CBE—Life Sciences Education*, *13*(2), 200–211.

Canada, K., & Pringle, R. (1995). The role of gender in college classroom interactions—a social-context approach. *Sociology of Education*, *68*(3), 161–186.

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.

Carter, A. J., Croft, A., Lukas, D., & Sandstrom, G. M. (2018). Women's visibility in academic seminars: Women ask fewer questions than men. *PLoS ONE*, *13*(9), e0212146.

Ceci, S. J., & Williams, W. M. (2010). Sex differences in math-intensive fields. *Current Directions in Psychological Science*, *19*(5), 275–279.

Cheryan, S., Siy, J. O., Vichayapai, M., Drury, B. J., & Kim, S. (2011). Do female and male role models who embody STEM stereotypes hinder women's anticipated success in STEM? *Social Psychological and Personality Science*, *2*(6), 656–664.

Cheryan, S., Ziegler, S. A., Montoya, A. K., & Jiang, L. (2017). Why are some STEM fields more gender balanced than others? *Psychological Bulletin*, *143*(1), 1–35.

Crawford, J. D. (1978). Career-development and career choice in pioneer and traditional women. *Journal of Vocational Behavior*, *12*(2), 129–139.

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.

Dasgupta, N. (2011). Ingroup experts and peers as social vaccines who inoculate the self-concept: The stereotype inoculation model. *Psychological Inquiry*, *22*(4), 231–246.

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.

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.

Finn, J. D., & Cox, D. (1992). Participation and withdrawal among fourth-grade pupils. *American Educational Research Journal*, *29*(1), 141–162.

Freeman, T. M., Anderman, L. H., & Jensen, J. M. (2007). Sense of belonging in college freshmen at the classroom and campus levels. *Journal of Experimental Education*, *75*(3), 203–220.

Fritschner, L. M. (2000). Inside the undergraduate college classroom—Faculty and students differ on the meaning of student participation. *Journal of Higher Education*, *71*(3), 342–+.

Gnesdilow, D., Evenstone, A., Rutledge, J., Sullivan, S., & Puntambekar, S. (2013). Group work in the science classroom: How gender composition may affect individual performance.

Goodenow, C., & Grady, K. E. (1993). The relationship of school belonging and friends' values to academic motivation among urban adolescent students. *Journal of Experimental Education*, *62*(1), 60–71.

Goulden, M., Mason, M. A., & Frasch, K. (2011). Keeping women in the science pipeline. *Annals of the American Academy of Political and Social Science*, *638*(1), 141–162.

Greenwald, A. G., & Krieger, L. H. (2006). Implicit bias: Scientific foundations. *California Law Review*, *94*(4), 945–967.

Haldane, T., Shehmar, M., Macdougall, C. F., Price-Forbes, A., Fraser, I., Petersen, S., & Peile, E. D. (2012). Predicting success in graduate entry medical students undertaking a graduate entry medical program. *Medical Teacher*, *34*(8), 659–664.

Haley, M. R., Johnson, M. F., & Kuennen, E. W. (2007). Student and professor gender effects in introductory business statistics. *Journal of Statistics Education*, *15*(3).

Harskamp, E., Ding, N., & Suhre, C. (2008). Group composition and its effect on female and male problem-solving in science education. *Educational Research*, *50*(4), 307–318.

Hoffmann, F., & Oreopoulos, P. (2009). A professor like me: The influence of instructor gender on college achievement. *Journal of Human Resources*, *44*(2), 479–494.

Jehn, K. A., Northcraft, G. B., & Neale, M. A. (1999). Why differences make a difference: A field study of diversity, conflict and performance in workgroups. *Administrative Science Quarterly*, *44*(4), 741–763.

Jensen, L., & Jensen, J. (1993). Family values, religiosity, and gender. *Psychological Reports*, *73*(2), 429–430.

Jones, S. M., & Dindia, K. (2004). A meta-analytic perspective on sex equity in the classroom. *Review of Educational Research*, *74*(4), 443–471.

Jurik, V., Groschner, A., & Seidel, T. (2013). How student characteristics affect girls' and boys' verbal engagement in physics instruction. *Learning and Instruction*, *23*, 33–42.

Kost, L. E., Pollock, S. J., & Finkelstein, N. D. (2009). Characterizing the gender gap in introductory physics. *Physical Review Special Topics—Physics Education Research*, *5*(1), 010101.

Kost-Smith, L. E., Pollock, S. J., & Finkelstein, N. D. (2010). Gender disparities in second-semester college physics: The incremental effects of a smog of bias. *Physical Review Special Topics—Physics Education Research*, *6*(2), 020112.

Kreutzer, K., & Boudreaux, A. (2012). Preliminary investigation of instructor effects on gender gap in introductory physics. *Physical Review Special Topics—Physics Education Research*, *8*(1), 010120.

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.

Leamaster, R. J., & Subramaniam, M. (2016). Career and/or motherhood? Gender and the LDS Church. *Sociological Perspectives*, *59*(4), 776–797.

Lee, M. (1993). Gender, group composition, and peer interaction in computer-based cooperative learning. *Journal of Educational Computing Research*, *9*(4), 549–577.

Lewis, K., Stout, J., Finkelstein, N., Pollock, S., Miyake, A., Cohen, G., & Ito, T. (2017). Fitting in to move forward: Using a belonging framework to understand gender disparities in persistence in the physical sciences, technology, engineering and mathematics (pSTEM). *Psychology of Women Quarterly*, *41*, 420–436.

Lewis, K. L., Stout, J. G., Pollock, S. J., Finkelstein, N. D., & Ito, T. A. (2016). Fitting in or opting out: A review of key social-psychological factors influencing a sense of belonging for women in physics. *Physical Review Physics Education Research*, *12*(2).

Lorenzo, M., Crouch, C. H., & Mazur, E. (2006). Reducing the gender gap in the physics classroom. *American Journal of Physics*, *74*(2), 118–122.

Luckenbill-Edds, L. (2002). The educational pipeline for women in biology: No longer leaking? *BioScience*, *52*(6), 513–521.

Ludlow, D. H. (1992). *Encyclopedia of Mormonism*. New York: Macmillan.

Madsen, A., McKagan, S. B., & Sayre, E. C. (2013). Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap? *Physical Review Special Topics—Physics Education Research*, *9*(2), 020121.

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.

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 USA*, *109*(41), 16474–16479.

Murphy, M. C., Steele, C. M., & Gross, J. J. (2007). Signaling threat: How situational cues affect women in math, science, and engineering settings. *Psychological Science*, *18*(10), 879–885.

Myaskovsky, L., Unikel, E., & Dew, M. A. (2005). Effects of gender diversity on performance and interpersonal behavior in small work groups. *Sex Roles*, *52*(9–10), 645–657.

Page, S. E. (2008). *The difference: How the power of diversity creates better groups, firms, schools, and societies* (new ed.). Princeton, NJ: Princeton University Press.

Pollock, S. J., Finkelstein, N. D., & Kost, L. E. (2007). Reducing the gender gap in the physics classroom: How sufficient is interactive engagement? *Physical Review Special Topics—Physics Education Research*, *3*(1), 010107.

Price, J. (2010). The effect of instructor race and gender on student persistence in STEM fields. *Economics of Education Review*, *29*(6), 901–910.

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.

Rich, Y., & Golan, R. (1992). Career plans for male-dominated occupations among female seniors in religious and secular high-schools. *Adolescence*, *27*(105), 73–86.

Rocca, K. A. (2010). Student participation in the college classroom: An extended multidisciplinary literature review. *Communication Education*, *59*(2), 185–213.

Schroeder, J., Dugdale, H. L., Radersma, R., Hinsch, M., Buehler, D. M., Saul, J., … & Johnson, P. J. (2013). Fewer invited talks by women in evolutionary biology symposia. *Journal of Evolutionary Biology*, *26*(9), 2063–2069.

Sekaquaptewa, D., & Thompson, M. (2002). The differential effects of solo status on members of high- and low-status groups. *Personality and Social Psychology Bulletin*, *28*(5), 694–707.

Shapiro, J. R., & Williams, A. M. (2012). The role of stereotype threats in undermining girls' and women's performance and interest in STEM fields. *Sex Roles*, *66*(3–4), 175–183.

Shaw, A. K., & Stanton, D. E. (2012). Leaks in the pipeline: Separating demographic inertia from ongoing gender differences in academia. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1743), 3736–3741.

Sheltzer, J. M., & Smith, J. C. (2014). Elite male faculty in the life sciences employ fewer women. *Proceedings of the National Academy of Sciences USA*, *111*(28), 10107–10112.

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.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, *52*(6), 613.

Steele, J., & Barling, J. (1996). Influence of maternal gender-role beliefs and role satisfaction on daughters' vocational interests. *Sex Roles*, *34*(9–10), 637–648.

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.

Tinto, V. (1997). Classrooms as communities—Exploring the educational character of student persistence. *Journal of Higher Education*, *68*(6), 599–&.

Walton, G. M., Cohen, G. L., Cwir, D., & Spencer, S. J. (2012). Mere belonging: The power of social connections. *Journal of Personality and Social Psychology*, *102*(3), 513.

West, J. D., Jacquet, J., King, M. M., Correll, S. J., & Bergstrom, C. T. (2013). The role of gender in scholarly authorship. *PLoS ONE*, *8*(7), e66212.

Wickware, P. (1997). Along the leaky pipeline. *Nature*, *390*(6656), 202.

Willoughby, S. D., & Metz, A. (2009). Exploring gender differences with different gain calculations in astronomy and biology. *American Journal of Physics*, *77*(7), 651–657.

Wilson, D., Jones, D., Bocell, F., Crawford, J., Kim, M. J., Veilleux, N., … & Plett, M. (2015). Belonging and academic engagement among undergraduate STEM students: A multi-institutional study. *Research in Higher Education*, *56*(7), 750–776.

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.

Young, D. M., Rudman, L. A., Buettner, H. M., & McLean, M. C. (2013). The influence of female role models on women's implicit science cognitions. *Psychology of Women Quarterly*, *37*(3), 283–292.

Zhan, Z., Fong, P. S., Mei, H., & Liang, T. (2015). Effects of gender grouping on students' group performance, individual achievements and attitudes in computer-supported collaborative learning. *Computers in Human Behavior*, *48*, 587–596.