CBE—Life Sciences Education
Vol. 13, 478–492, Fall 2014

Article

Gender Gaps in Achievement and Participation in Multiple Introductory Biology Classrooms

Sarah L. Eddy,*† Sara E. Brownell,†‡ and Mary Pat Wenderoth*

*Department of Biology, University of Washington, Seattle, WA 98195; ‡School of Life Sciences, Arizona State University, Tempe, AZ 85287

Submitted October 28, 2013; Revised May 20, 2014; Accepted May 23, 2014
Monitoring Editor: Deborah Allen

Although gender gaps have been a major concern in male-dominated science, technology, engineering, and mathematics disciplines such as physics and engineering, the numerical dominance of female students in biology has supported the assumption that gender disparities do not exist at the undergraduate level in life sciences. Using data from 23 large introductory biology classes for majors, we examine two measures of gender disparity in biology: academic achievement and participation in whole-class discussions. We found that females consistently underperform on exams compared with males with similar overall college grade point averages. In addition, although females on average represent 60% of the students in these courses, their voices make up less than 40% of those heard responding to instructor-posed questions to the class, one of the most common ways of engaging students in large lectures. Based on these data, we propose that, despite numerical dominance of females, gender disparities remain an issue in introductory biology classrooms. For student retention and achievement in biology to be truly merit based, we need to develop strategies to equalize the opportunities for students of different genders to practice the skills they need to excel.

INTRODUCTION

Women are underrepresented in undergraduate science, technology, engineering, and mathematics (STEM) majors (National Science Foundation [NSF], 2011). Even fewer women pursue graduate school and careers in STEM fields, particularly careers in academia (Handelsman, 2005; National Research Council [NRC], 2007; Beede *et al*., 2011; NSF, 2011). The possible reasons for the gap in the persistence of females compared with males in STEM, frequently referred to as the "leaky pipeline," are numerous and multifaceted (Clark Blickenstaff, 2005; Burke and Mattis, 2007), and despite a concentrated effort by funding agencies directed at both K–12 and colleges, the problem persists.

DOI: 10.1187/cbe.13-10-0204
Address correspondence to: Sarah L. Eddy (sleddy@u.washington.edu).
†These authors contributed equally to this work.

The one exception to this pattern of underrepresentation of females in STEM is the field of biology. Women account for more than 60% of undergraduate biology majors and approximately half of all graduate students in the biosciences (Luckenbill-Edds, 2002; Amelink, 2009), unlike other STEM disciplines such as physical sciences, in which women make up only 43% of undergraduates (Amelink, 2009) and 20% of graduate students (Mulvey and Nicholson, 2011, 2012). Owing to the significant numbers of females pursuing biology, it is often assumed that biology is a STEM discipline that has overcome gender[1] disparities. In fact, this assumption is so prevalent that studies in chemistry and physics sometimes use biology as a positive control for comparisons of the observed gender gaps in their fields (e.g., Ferreira, 2003; Ecklund *et al*., 2012).

Gender inequity in biology does emerge at the postgraduate level, however, as fewer female biologists pursue postdoctoral work or positions in academia relative to males

[1]Gender is a complicated identity based on a person's internal experience of who he or she is, not the sex he or she was assigned at birth (which is determined by physical, hormonal, or chromosomal characteristics). For example, a person can be assigned female at birth (sex), but identify as male or as neither male nor female (gender). Many education studies, including ours, use self-reported demographic information, which is a measure of gender rather than sex.

(NSF, 2011). In addition, the observed distribution of professional prestige demonstrates gender inequities: even in a female-dominated field, women are less likely to be selected to participate in symposia, especially if they are organized by men (Isbell *et al*., 2012), and women are less likely to be first authors on biology papers as compared with other more "caring" fields (Lariviere *et al*., 2013). Published explanations for these differences are often based on individual-level decisions such as work/life balance preferences or the desire to start a family (Ceci and Williams, 2010, 2011; Rosser 2012), as opposed to systematic institutional challenges.

Although gender disparities in biology have been primarily documented at the graduate level and later academic life, it is likely that student experiences at the K–16 levels influence these later outcomes. Exploring the potential for gender disparities early in a student's college experience seems particularly important, because early experiences and decisions, such as choosing a major or being recognized as competent by a biologist, are the first steps in the process of developing a professional biology identity (Cech *et al*., 2011). Although it is important to track the retention of students, this type of coarse-grained measure does not provide insights into the underlying mechanisms that may impact the experiences of female students. In studies in education, psychology, and sociology, measures of the experience of students that have been shown to be related to retention include: academic achievement (Carrell *et al*., 2010; Kost-Smith *et al*., 2010), interest in the discipline (Kost-Smith *et al*., 2010), class participation (Holmes, 1992; Guzzetti and Williams, 1996), science identity (Meece *et al*., 2006), professional role confidence (Cech *et al*., 2011), access to resources (Jovanovic and King, 1998), self-efficacy (Meece *et al*., 2006), and course-related anxiety (Pomerantz *et al*., 2002). In our study, we focus on two measures that an instructor could easily document in his/her own course: exam achievement and participation in whole-class discussions.

The most commonly studied gap in STEM fields is achievement, which is a strong predictor of retention in STEM disciplines, particularly relative to achievement in non-STEM courses (Beasley and Fischer, 2012; Riegle-Crumb *et al*., 2012). Studies conducted in biology and biochemistry on gender differences in achievement offer conflicting results. Rauschenberger and Sweeder (2010) showed that females underperformed compared with males in an introductory-level biochemistry course when prior ability was controlled for, and a subsequent study showed that females also systematically scored lower than males in upper-division biology courses (Creech and Sweeder, 2012). However, Strenta (1994) failed to find a gender gap in persistence in biology or any STEM field after controlling for student ability, although there was a trend for gender differences in confidence about ability and feelings of depression about progress, with females having lower confidence and more feelings of depression. More recently, Lauer *et al*. (2013) saw no difference in academic achievement between males and females in two courses: introductory biology and biochemistry. These conflicting results indicate that the question of gender gaps in academic performance at the introductory biology level remains unresolved.

A second potential gender gap that could occur in biology classrooms is a gap in participation. Although, to our knowledge, prior research has not been done on participation in undergraduate biology classrooms, there is a large body of literature on participation across disciplines at the college level. It is evident that instructors at the college level value classroom participation; instructors dedicate 2–23% of class time to student participation in a typical lecture class (Nunn, 1996), and this percentage can be even higher in active-learning classrooms (Smith *et al*., 2013). In addition, a student's greater level of participation in class has been linked with positive perception of the class (Crombie *et al*., 2003), decreased anxiety about performance and ability in the course topic (Fassinger, 2000), and increased critical thinking (Tsui, 2002). In studies at the college level, the pattern of participation by men and women in whole-class discussions is not consistent, with various studies showing a bias in either direction (more female than male participation: Howard and Henney, 1998; Fritschner, 2000; Howard *et al*., 2006; more male than female participation: Crombie *et al*., 2003; Tatum *et al*., 2013; or no difference: Cornelius *et al*., 1990; Pearson and West, 1991; Brady and Eisler, 1999). Although many classrooms were observed in these studies, none of them were courses in STEM disciplines nor were they conducted in classrooms the size of typical large-enrollment introductory-level STEM courses.

Two possible explanations for the conflicting achievement and participation patterns at the undergraduate level include: 1) variation in student populations and 2) variation in the instructors. When discussing the experience of female students in a discipline, it is important to recognize that students are not monolithic. Gender is a complicated identity based on a person's internal experience of who he or she is. Thus, individuals can vary in the degree to which they identify with their gender, the gender roles associated with their gender, and how their gender identity influences their experience in different settings such as a classroom (Nosek *et al*., 2002; Schmader *et al*., 2004; Lane *et al*., 2012). In addition, gender is only one of a multitude of social identities that make up who we are and how we react in certain settings. Other identities, such as a student's race/ethnicity, could modify a student's experience as a female in a classroom (Ong *et al*., 2011), and there have been a few studies that have looked at the interactions between race/ethnicity and gender (Anderson, 2005; Riegle-Crumb *et al*., 2011).

Just as all females are not the same, not all biology classrooms are the same. The classroom experience can be influenced by a multitude of factors, including teaching methods (Beichner *et al*., 2007), who is enrolled in the class (Theobald and Freeman, 2014), and whether the course is optional or required (Brownell *et al*., 2013). One classroom factor that has been found to have a specific influence on achievement and participation is instructor gender. In STEM courses, females participated more and identified more with the subject matter when instructors were female (Stout *et al*., 2011; Young *et al*., 2013). Some studies have found that instructors of the same gender, particularly instructors students perceive as competent, can improve the performance of female students (Haley *et al*., 2007; Hoffman and Oreopoulos, 2009; Carrell *et al*., 2010; Antecol *et al*., 2012), while other studies found no difference (Griffith, 2010; Price, 2010; Stout *et al*., 2011). Thus, an instructor effect in college-level STEM courses remains a contentious issue that would benefit from further exploration.

In our large, retrospective study, we tested the hypothesis that there are no gender disparities in undergraduate biology for achievement or whole-class participation, using a data

set that included 23 classes, 26 instructors, and almost 5000 students across three large introductory biology courses for majors at a large research institution. We explicitly tested two additional subhypotheses: 1) whether instructor gender influences these patterns of female participation and achievement and 2) whether student racial/ethnic identity modifies the relationship between student gender and achievement.

METHODS AND RESULTS

The Classes

This study examined 23 individual offerings (called classes) of the three courses composing the introductory biology series for science majors over a recent 2-yr period at a large public R1 university on the quarter system. The first course in the series focuses on evolution and ecology; the second on molecular, cellular, and developmental biology; and the third on plant and animal physiology. Students taking the introductory biology series are predominately sophomores and biology majors. Although this is a three-course series, not all science majors are required to take all three. Individual classes ranged in size from 159 to more than 900 students, depending on the term. Teaching methods varied between instructors; some were taught with exclusively passive teaching methods, while others were highly student centered and interactive. In addition, exam format varied from almost exclusively essay questions to exclusively multiple choice, with the majority of classes using short-answer exam formats.

Although some classes were taught by one instructor (33.3%), most classes were cotaught by two instructors (66.7%), each teaching for 5 wk. In total, 26 different instructors taught these 23 classes. Instructor gender also varied across these classes: 33.3% were taught exclusively by either one or two male instructors, 37.5% had both a male and a female instructor, and 29.2% had either one or two female instructors.

During the 2-yr period, more than 5000 students enrolled in the series. Demographic information collected by the university registrar revealed that on average 58.1% of the students in these classes identified as female, but this number ranged from 53 to 64%, depending on the specific class. In addition, on average 43.2% of students identified as white, 37.6% Asian, 2.5% black, 0.8% Hawaiian and Pacific Islander, 4.8% Latin@[2], 1.1% Native American, and 3.4% did not identify a racial or ethnic group. An additional 6.6% were international students.

Study 1: Is There a Gender Achievement Gap in Introductory Biology?

Methods. We collected student exam performance for the 23 introductory biology classes along with the following demographic data obtained from the registrar: student gender identity (0 = male, 1 = female), student racial/ethnic/national identity (Asian, Black, Hawaiian/Pacific Islander, Latin@, Native American, White, and International), and college grade point average (GPA) upon entry into the introductory biology series. We also recorded classroom-level variation in the gender identity of the instructors as 0 = no female instructors, 1 = half of class taught by a female instructor, 2 = all of class taught by female instructor(s).

The response variable for our analysis was overall performance on exams in the class. The number of points allocated to exams varied from class to class (from 200 to 400 points), but because the focus of this project was to compare the relative position of males and females within a classroom and not to document the absolute value of their performance in the class, we standardized exam scores by transforming them into *z*-scores based on each classroom's mean and SD. On average, each student was represented in our analysis twice.

Accounting for Differences between Students. Students vary in many ways that could influence exam performance. We hypothesized that exam performance will be influenced by gender and ethnicity and therefore included those terms in our analyses. In addition, because this study, like many educational studies, has a quasi-random design (due to students selecting into classes rather than being randomly assigned to classes), it is possible for inherent differences in students outside of our variables of interest to potentially bias results (Theobald and Freeman, 2014). To limit this possibility, we include two kinds of covariates in our analyses that can account for potential differences between students: 1) a measure of student performance in college and 2) a random-effect term that captures between-student variation (see *Statistical Analyses* below). To account for potential covariation between academic preparedness and our response variables (exam performance and participation), we include cumulative college GPA as a covariate in our analyses, as it has been shown to be a strong predictor of student achievement in a number of previous studies (Xie and Shauman, 2003; Freeman *et al.*, 2007a, 2011; Riegle-Crumb and King, 2010; Haak *et al.*, 2011; Creech and Sweeder, 2012). In addition, including a covariate that captures some aspects of academic preparedness in our models allows us to more rigorously test the effect of our variables of interest (e.g., student gender and ethnic identity) on our outcome variable (e.g., student exam performance; Lipsey *et al.*, 2012). Therefore, when we describe our results, we are describing differences in exam performance between males and females with the same cumulative college GPAs at entry into the introductory biology series.

Accounting for Differences among Classrooms. As instructional practices (Lorenzo *et al.*, 2006; Beichner *et al.*, 2007) and exam difficulty/format (Dimitrov, 1999; Bell and Gafni, 2000) have been shown to influence the relative achievement of males and females, we include a variable for class as a random effect in our analyses (see *Statistical Analyses* below). This random effect captures the variation in performance between classes that is not related to our predictor variables and allows us to compare outcomes across different class and exam structures. The strength of this approach is that we can control for individual class variation that could be due to anything that may be different among the courses (e.g., the instructor, the

[2]Latin@ is a gender inclusive way of describing people of Latin American descent (Demby, 2013). The term is being increasingly used in the Latin@ community including national organizations such as the National Latin@ Network and college groups such as Movimento Estudiantil Chican@ de Azltan. Academic departments across the country are also recognizing and adopting this term. For example, the University of Wisconsin, Madison, as well as the University of San Francisco now have Chican@/Latin@ Studies Programs.

students, exam format, exam difficulty, or something that we have not even considered).

Statistical Analyses. Similar to many educational studies using multiple classes or schools, the data in this study are hierarchically nested. As students are nested in classes, we have explanatory variables at both the student level (student gender identity and cumulative college GPA) and the class level (instructor gender identity and term and course). The hierarchical nature of this data set is important to account for, because a student's exam performance is likely to be more similar to a classmate's performance than a student outside of his or her class, as students in the same class share the same exams and instructional environment (Kreft and de Leeuw, 2002). In cases like this, linear regression can lead to erroneous conclusions, because the assumption of independence of observations is violated (i.e., 100 students in the same class are not 100 independent data points; Kreft and de Leeuw, 2002). A statistical method called *multilevel modeling* has been developed to account for this nonindependence of data in nested-data structures and is widely used in the fields of education, sociology, and ecology (Paterson and Goldstein, 1991; Kreft and de Leeuw, 2002; Raudenbush and Bryk, 2002; Bolker *et al*., 2009).

Multilevel models differ from traditional linear regression models in many ways. First multilevel models are a type of mixed-effects model that includes fixed and random effects. Fixed effects are generally the variables of interest, and, in linear regressions, all variables are assumed to be fixed. In mixed-effects models, some variables are allowed to be random effects. Random effects are those that can be seen to be drawn at random from a population. This allows for inference beyond the specific populations measured (Kreft and de Leeuw, 2002). For example, the particular classes students are enrolled in could be considered a random effect if the subset of classes used in a study can be seen as having been chosen at random from a larger pool of possible classes. Random effects are also the variables that can account for clustering in a nested-data structure (Bolker *et al*., 2009). A second way multilevel models differ from linear models is the ability to account for interactions that occur across levels of the hierarchy. For example, its possible that the relationship between Scholastic Aptitude Test (SAT) verbal score and exam performance might differ for a student in a class with multiple-choice exams versus essay exams. Multilevel models can account for this by incorporating a random slope for SAT verbal (i.e., allowing the slope of the relationship between SAT verbal and *z*-score to vary from class to class).

In our model, the outcome variable was *z*-scores from exam performance. Student gender identity, student racial/ethnic identity, college GPA, and instructor gender identity were fixed effects, and class and student were random effects. Having student identification as a random effect allows us to account for repeated measures on the same student and avoids issues of pseudoreplication. In preliminary analyses used to develop a baseline model (cf. Zuur *et al*., 2009), we found that only the relationship between college GPA and *z*-score varied from class to class, so we used a random slope model that allowed the slope of the regression line to vary by class to account for these differences. These preliminary results indicate that the size of the gender gap is not a unique feature of a particular combination of course structure, exam format, or instructor. In this study the only class-level factor we were able to isolate was instructor gender identity. It may be possible with data collection beyond the scope of this work to parse out the impact of specific exam formats and/or different instructional practices on student achievement, and these are potential areas of future research. Multilevel models were analyzed in R using the lme4 package (Bates *et al*., 2013).

To identify which fixed-effect variables best explained the patterns in student exam scores, we used a powerful multimodel inference technique using Akaike's information criterion (AIC; Akaike, 1973). This statistical method is commonly used in the fields of ecology, evolution, and behavior when data come from observational studies with a large number of possible explanatory variables. It has begun to be used in educational studies focused on large student populations that have a large number of possible explanatory variables (Haak *et al*., 2011). Several authors have argued that multimodel inference is a more rigorous approach to model selection and variable selection in regression analyses than the more common method of simple significance testing (Akaike, 1974; McQuarrie and Tsai, 1998; Anderson *et al*., 2000; Johnson and Omland, 2004; Burnham *et al*., 2011; Garamszegi, 2011; Symonds and Moussalli, 2011). In addition, this type of multimodel inference method avoids some of the common issues of stepwise model-selection methods, including the inconsistencies in model selection that result from different stepwise methods and criteria (reviewed in Hegyi and Garamszegi, 2011).

We used this multimodel inference technique using the Akaike's information criterion corrected for small sample sizes (AICc) on linear mixed-effects regression models with a continuous outcome variable: student exam performance (*z*-scores). AICc estimates the likelihood that each possible model is the best model given our sample size (Akaike, 1973; Anderson, 2008). These AICc values are then used to rank the models. From these AICc values, AICc differences (Δi) and Akaike weights (ω_i) are calculated. The Δi represents the strength of evidence in support of each model as the best model. The larger the Δi, the less likely the model. Models with an $\Delta i > 10$ are considered poor predictors compared with the best model and thus are not included in our analyses (Burnham and Anderson, 2004). Akaike weights (ω_i) are a calculation of the likelihood of the observed data given a particular model that has been standardized so the sum of all the model weights add up to one. These weights make it easier to compare models, as the likelihood is approximately the probability that the model is actually the best model. AICc analyses were implemented in R using the MuMIn package (Barton, 2013).

In addition to identifying the best model, the multimodel inference approach also allows us to use information from all possible models to generate regression coefficients through model averaging (Anderson, 2008; Garamszegi, 2011). This method of calculating regression coefficients accounts for the underlying uncertainty that is always present as to which model best fits the data. Akaike weights can also be used to calculate a measure of the relative importance of an explanatory variable (Anderson, 2008; Garamszegi, 2011). This process involves summing the Akaike weights across all the models that include a particular explanatory variable. This relative variable importance is the probability that a

particular variable is important for explaining observed differences in exam performance.

Six potential fixed variables were initially considered to contribute to student exam performance (Z.Score): 1) cumulative college GPA upon entry into the biology series (Cum.GPA); 2) student gender identity (a factor with two levels; Stu.Gender); 3) student race/ethnicity/nationality (a factor with seven levels; Ethn); 4) an interaction between student gender identity and race/ethnicity/nationality (Stu.Gender*Ethn); 5) instructor gender identity (a factor with three levels; Inst.Gender); and 6) an interaction between student gender identity and instructor gender identity (Stu.Gender*Inst.Gender). Only students with a complete set of these variables were included in this analysis. Combinations of these variables produced a total of 26 potential models to describe our data. The total number of models tested was substantially lower than our number of observations ($n = 7841$ students), which justified fully exploring this set of models. Thus, we systematically explored the possible models for our data and ultimately chose the model that best fits the data according to the model-selection statistics. We also calculated the model averaged regression coefficients for the fixed effects in our model. Our initial full model was as follows:

$$\begin{aligned} \text{Z.Score} = {} & \text{Cum.GPA} + \text{Stu.Gender} + \text{Ethn} \\ & + \text{Stu.Gender}^{*}\text{Ethn} + \text{Inst.Gender} \\ & + \text{Stu.Gender}^{*}\text{Inst.Gender} \\ & + (1|\text{Stu.ID}) + (\text{Cum.GPA}|\text{Class}) \end{aligned} \quad (1)$$

This model includes the random terms for student identity (represented by 1|Stu.ID) and the interaction between cumulative GPA and class (represented by Cum.GPA|class).

Results for Study 1: Is There a Gender Achievement Gap in Introductory Biology? Using model selection, we found six models with reasonable support ($\Delta i < 10$) that explained the patterns in exam performance across the 23 classes (Table 1). The top two models had the majority of the support (summed $\omega = 0.71$). The best model included three of the six possible fixed effects (cumulative college GPA, student gender identity, student race/ethnicity/nationality). The second-best model included the two instructor variables (instructor gender and student gender identity*instructor gender).

The main effect of identifying as female across all our models was to decrease exam performance by ~0.2 of an SD ($\beta = 0.21 \pm 0.04$ [SE], p value < 0.0001; Table 2). The student gender identity variable had a relative variable importance of 1 and was present in all of the six best models, implying that gender had a consistent and reliable impact. That is, if two students are in the same class and have the same GPA and race/ethnicity/nationality, but one student is male and the other is female, our model predicts that the female student will score 0.2 SDs lower in the distribution of scores in the class. In classes with 400 exam points ($n = 19$), the average SD was 42.8 points. Thus, female students are scoring, on average, 11 points (2.8%) lower on their overall exam grades than male students with the same GPA.

A main effect of race on exam points was well supported in our analyses (relative variable importance = 1, present in all six of the best models). However, the interaction between student gender identity and race/ethnicity/nationality was not supported; the interaction term had the lowest relative variable importance (0.18) of all the predictors included in the model (Table 2). It also is present only in the fifth most well-supported model, and this model does not have much support relative to the best model ($\omega_i = 0.07$ out of 1; Table 1).

The model averaged coefficients reveal that the only significant interactions between gender identity and racial/ethnic identity is within Latin@s (0.22 ± 0.1, p value $= 0.026$; Table 2). In a class taught exclusively by males, if a white male student, a Latina student, and a Latino student enter that class with the same cumulative college GPA, the Latina student is predicted to perform 0.23 SDs lower than the white male (-0.24*Latin@ + -0.21*StudentGenderF + 0.22*Latin@*StudentGenderF) and the Latino student is predicted to perform 0.24 SDs lower than the white male (-0.24*Latin@). Thus, although there is no difference between being male and female for Latin@s, both underperform compared with white males. This lack of gender gap for Latin@ students could be attributed to males experiencing a great cost for being Latino in the introductory biology classroom compared with females, or it could be attributed to Latinas experiencing less of a cost of being female than other racial groups. It is impossible to distinguish between these hypotheses from this type of observational data, but it could be interesting to further explore the experience of Latin@ students, as this pattern is unique among the racial and ethnic groups in this study.

Table 1. Best models include student gender identity as a predictor of exam performance[a]

Rank	Model[b]	AICc	Δi	ω_i
1	Cum.GPA + Ethn + Stu.Gender	18019.9	0	0.41
2	Cum.GPA + Ethn + Stu.Gender + Inst.Gender + Stu.Gender*Inst.Gender	18020.6	0.63	0.30
3	Cum.GPA + Ethn + Stu.Gender + Inst.Gender	18022.5	2.58	0.11
4	Cum.GPA + Ethn + Stu.Gender + Ethn	18023.1	3.21	0.08
5	Cum.GPA + Ethn + Stu.Gender + Stu.Gender*Ethn + Inst.Gender + Stu.Gender*Inst.Gender	18023.5	1.95	0.07
6	Cum.GPA + Ethn + Stu.Gender + Inst.Gender	18025.7	5.79	0.02

[a]Relative ranking (from most support to least) of six best models for predicting student exam performance using AICc model selection. Only models that are informative ($\Delta i < 10$) are shown. The table shows only fixed-effect terms, but all models also include two random-effect terms: Student and an interaction between cumulative college GPA and the class students were enrolled in.

[b]Cum.GPA = cumulative college GPA at start of introductory biology series; Stu.Gender = student's gender identity; Ethn = student ethnic/racial/national identity; Inst.Gender = instructor(s) gender.

Table 2. Female gender significantly decreases exam performance relative to males across 23 introductory biology classes[a]

Parameter	Relative variable importance	Model averaged regression coefficient ± SE	*p* Value[b]
Intercept	NA	−4.10 ± 0.20	<**0.0001**
Student-level variables:			
Cumulative GPA	1	1.32 ± 0.06	<**0.0001**
Ethnicity/Race/Nationality:	1		
(reference level: White)			
Asian		−0.13 ± 0.03	<**0.0001**
Black		−0.43 ± 0.09	<**0.0001**
Hawaiian/Pacific Islander		−0.22 ± 0.14	0.114
International		−0.44 ± 0.06	<**0.0001**
Latin@		−0.24 ± 0.07	**0.001**
Native American		−0.24 ± 0.11	**0.030**
Student Gender:	1		
(reference level: Male)			
Female		−0.21 ± 0.04	<**0.0001**
*Ethnicity/Race/Nationality*Student Gender:*	0.18		
*(reference levels: White*Male)*			
Asian*Female		−0.01 ± 0.05	0.830
Black*Female		0.17 ± 0.14	0.227
Hawaiian/Pacific Islander*Female		0.19 ± 0.23	0.412
International*Female		−0.08 ± 0.09	0.383
Latin@*Female		0.22 ± 0.10	**0.026**
Native American*Female		0.13 ± 0.19	0.492
Classroom-level variables:			
Instructor Gender:	0.51		
(reference level: Only Male)		−0.08 ± 0.07	0.27
1 Female/1 Male			
Only Female		−0.01 ± 0.08	0.90
*Student Gender*Instructor Gender:*	0.37		
*(reference levels: Male*Only Male)*		0.07 ± 0.04	0.055
Female Student*1 Female/1 Male Instructor			
Female Student*Only Female Instructor(s)		0.10 ± 0.05	0.024

[a]Model-averaged regression coefficients and relative variable importance for all six possible fixed-effect terms. Although not shown, this model includes two random-effect terms: (1|Stu.ID) + (Cum.GPA|class).

[b]Bolded *p* values are significant.

Instructor gender and the interaction between instructor gender and student gender were present in the second-best model (Table 1) and have relative variable importances of 0.5 and 0.37, respectively (Table 2). This indicated that there is more uncertainty about their importance than the variable of student gender identity, student race/ethnicity/nationality, and cumulative college GPA. Using the model average coefficients that incorporate this uncertainty about the relationship between instructor gender identity and student performance, we find it is only the interaction between student gender identity and females exclusively teaching the class that has a significant positive impact on student exam performance ($\beta = 0.10 \pm 0.05$, $p = 0.024$; Table 2). Thus, if a course was taught solely by female instructors, the achievement gap between students of different genders with the same cumulative college GPA and race/ethnicity/nationality would be reduced by 62%. This would mean the gender gap in a class with two female instructors would be reduced from 11 points (a gap of 2.8%) to 7 points (a gap of 1.7%).

Study 2: Are There Gender Gaps in Participation during Whole-Class Student–Instructor Interactions?

Methods. Over the 2-yr period, 26 instructors taught the introductory biology series. Though many instructors taught the courses more than once during this 2-yr period, participation data were only collected from one quarter for each of the 26 instructors. We observed individual class sessions to determine participation rates. Kane and Staiger (2012) found that two trained individuals each observing a single 45-min session of a teacher's class have a reliability score of 0.67 (i.e., observations are more likely to be due to a characteristic of the teacher and not due to a particular observer), and this paired observation of one session is just as reliable as having independent observations of four sessions. In our study, to be conservative and to increase the number of student–teacher interactions sampled, we randomly selected three class sessions for each instructor. These 78 videos were scored by two observers, one male and one female, who recorded 1) the ways in which students verbally interact with the instructor during class and 2) the perceived gender of any student who spoke out during class. In this study, we focused solely on student verbal interactions that occurred in the context of the whole class. Although there are other ways for students to interact in class (e.g., asking an instructor a question during small-group work), it was impossible for us to analyze those conversations through our whole-class video recordings.

We categorized student interactions in front of the whole class in the following ways: 1) asking a spontaneous question, 2) volunteering to answer an instructor-generated question,

or 3) responding to an instructor-generated question when called on by the instructor through random call. An event was coded as a spontaneous student question when a student asked an instructor an unprompted question or was only very generally prompted: "Does anyone have a question?" Volunteer responses were characterized by students raising their hands or shouting out answers of their own volition in response to instructor questions. In these volunteer responses, only those students who choose to participate did. Random call required students to be more accountable for participating in class than either of the two previous methods. Random call has a particular structure that is similar to cold-calling, with the instructor calling on students by name to answer questions the whole class hears. However, random call differs from cold-calling in that an instructor does not decide upon whom he or she will call. Instead, an instructor comes to class with a randomized class list and calls student names in the order the names appear on this list. Observers were able to distinguish random call from volunteer responses in the videos by watching the instructor behaviors. In random call, the instructor calls out student first and last names without waiting for volunteers and can often be seen referring to a list before saying a student name.

Only instructors who had a total of five or more students participate in any one of these three types of student–instructor interactions across the three observed class sessions were included in the analysis. We chose five as a lower cutoff to be conservative, as the analysis we planned to use involved ratios. With ratios, the fewer observations, the easier it is to see extreme values that would be classified as significant deviations from expected. Based on this criterion, only 20 of the 26 instructors qualified for analysis of student participation in whole-class interactions.

The two observers also independently assigned a gender to the participating students in the videos based on the students' visual appearances and/or auditory characteristics. If observers could not identify the gender of the speaker or did not agree on the gender, the student was marked as "cannot determine." Overall, observers could not assign a gender to 7.9% of the students who spoke in front of the whole class. If more than 20% of the total number of students speaking in the three sessions could not be assigned a perceived gender, then the instructor teaching that class was not included in our analysis. This occurred only for two instructors in which either the camera was too far away to see any of the students who spoke or students spoke so briefly it was impossible to identify them. Therefore, of the 20 instructors who had a total of more than five students speak out to the whole class over three class sessions, we were only able to analyze participation data for 18 instructors.

We chose to work with historic video data so that we did not influence instructor behavior by sitting in and recording real-time interactions. However, the methods used in this study have several limitations. The first disadvantage of working with historic video data is that we cannot identify individual students by name in order to determine their self-reported gender identity. Perceived gender was the best proxy we could collect, but perceived gender does not always align with self-identified gender (e.g., a male student with long hair may be mistakenly identified as a female student, or a student who appears to be female based on physical characteristics may actually self-identify as male). Second, in the majority of our observed classrooms, an individual instructor used multiple student-engagement techniques (volunteers and student questions) as well as small-group work. Thus, it was not possible for us to link exam performance (i.e., academic achievement) in these classes to interaction methods used, because multiple methods were used, and it was impossible to ascertain the independent impact of one of these methods on exam performance.

Statistical Analyses. Analyses were run separately for each type of student–instructor interaction (spontaneous questions, volunteer discussions, and random call) to determine whether there are gendered patterns of participation under each strategy. Some of the instructors ($n = 4$) had enough student participants in two categories to be included in both sets of analyses, and a few ($n = 2$) exceeded the minimum number of students for all three methods. Therefore, an individual instructor could be included in the analysis of more than one type of interaction. Overall, 11 instructors were included in the analysis for spontaneous student questions, 13 in the analysis of volunteer-based discussions and four in the analysis of random call discussions. As the number of student–instructor interactions varied widely between these 18 instructors, results will be expressed as percentage of interactions by females. Because only a small number of students were in each instructor analysis, an exact binomial test for goodness of fit was used to compare the expected value of female speakers (the percentage of women enrolled in the class) with the observed percentage of female voices heard in each interaction type. To explore the gender bias in each interaction type across all instructors, a two-tailed t test was performed across all the instructors for student questions, volunteer responses, and random call, individually. In addition a nonparametric Kruskal-Wallis analysis of variance was performed to determine whether instructor gender influenced female response rates. Analyses were implemented in R (R Core Team, 2012).

Results for Study 2: Are There Gender Gaps in Participation during Whole-Class Discussions? Across 11 classrooms that had spontaneous student questions, there was not a significant difference (two tailed t test: $p = 0.319$) between the proportion of females enrolled in a class ($58.7 \pm 3.5\%$ SD) and the proportion of questions asked by females ($39.9 \pm 22.5\%$). Although the summary t test did not reveal a significant difference across the 11 classes, the exact binomial tests within each class identified five classrooms in which females asked fewer questions than expected ($p < 0.03$) and six classrooms for which there was no statistical difference (Figure 1). In no classrooms did females significantly ask more questions than males.

On the other hand, across the 13 classrooms in which there were volunteer responses, the number of responses attributed to females ($36.7 \pm 12.9\%$) was significantly lower ($p = 0.042$) than would be expected based on the number of females enrolled in each class ($59.2 \pm 3.6\%$). There was less variation from class to class in this result relative to the variation in spontaneous questions: nine of the 13 classrooms revealed significant differences ($p < 0.05$) between observed and expected number of female volunteer responses (Figure 2). In no classrooms were females heard more than males when the instructor solicited volunteer responses.

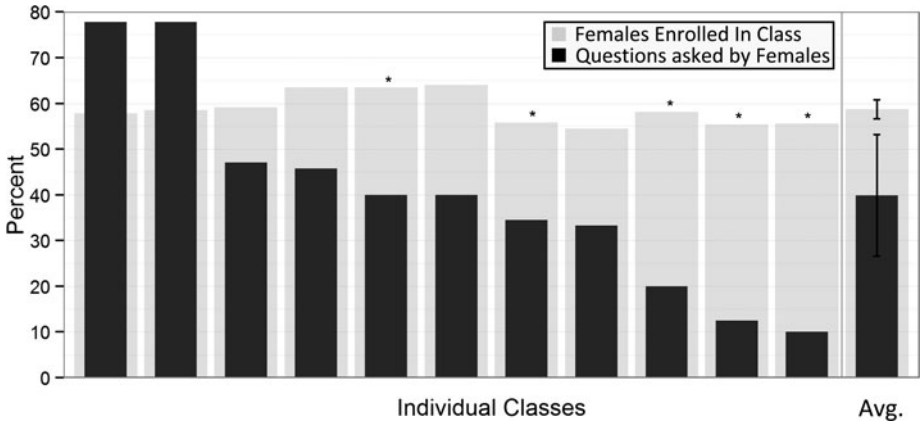

Figure 1. Variation by class in the percentage of questions asked by females. Comparison of the percentage of females in a class (gray bars) with percentage of unprompted questions in class asked by females (nested black bars). Asterisks (*) indicate that the exact binomial test was significant at the $p = 0.05$ level.

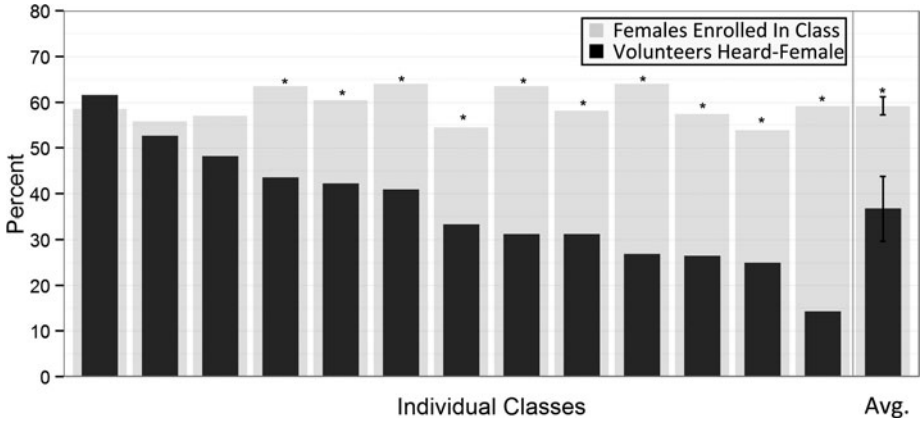

Figure 2. Females heard in volunteer student–instructor interactions significantly less than expected based on enrollment. Comparison of the percentage of females in a class (gray bars) with percentage of volunteer-based student–instructor interactions that involved female students (black bars). Asterisks (*) indicate that the exact binomial test was significant at the $p = 0.05$ level.

Unlike spontaneous student questions or volunteer responses, there were no significant gender differences in participation when participation was based on random call ($p = 0.9$). In this case 61.0 ± 0.04% of students in the class were female and 60.0 ± 11.8% of the participants in random call were female (Figure 3). This pattern was consistent across the four classrooms in which random call was used.

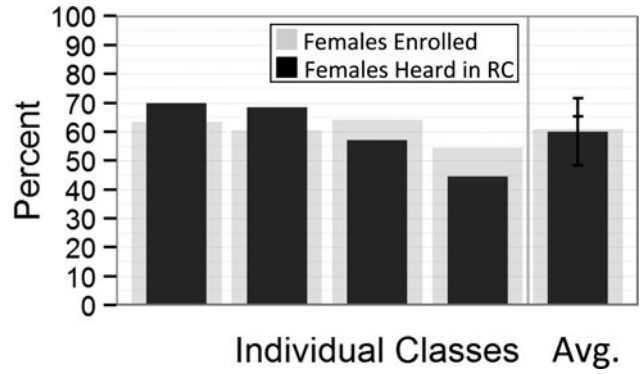

Figure 3. Random call extinguishes gender gap in whole-class participation. Comparison of the percentage of females in a class (gray bars) with percentage of females who are called on during random call (RC)-based discussions (nested black bars).

We found no evidence that instructor gender moderated any of these participation patterns (volunteer: $\chi 2 = 0.34$, $df = 1$, $p = 0.56$; student questions: $\chi 2 = 0$, $df = 1$, $p = 1$).

DISCUSSION

In our study of 23 classes at an R1 university, we found evidence of systematic gender-based gaps in both exam achievement and whole-class participation. Female students underperformed on exams compared with male peers with similar historical college performance. Furthermore, female voices were heard much less frequently than would be expected based on the gender composition of the classes. The causes and consequences of these subtle disparities are difficult to discern, but they could have lasting impacts on the development of a science identity, sense of belonging, and confidence of female science majors, which may have negative effects on long-term retention of women in the field of biology (Wickware, 1997; Johnson, 2007; Collett *et al.*, 2013).

Small, Yet Potentially Important Achievement Gap between Males and Females

In this study, we found that the exam performance of female students was consistently a quarter of a SD lower than the performance of male students with similar college GPAs, leading to an average 2.8% difference in exam scores. In

addition, the main effect of gender was significant even when an interaction between gender and race/ethnicity/nationality was added. This indicates that the impact of being female in a biology classroom is consistent across the different racial/ethnic groups present in the observed classrooms. If the main effect terms had been nonsignificant, but interaction terms between gender and race/ethnicity/nationality were significant, this would have indicated that gender only had a significant impact for some racial/ethnic groups. This was not the pattern we observed: the only group with a significant interaction term between race/ethnicity and gender was Latin@s. Replication of this difference and more detailed studies will be necessary to parse out the significance of the difference between the experience of Latin@s in the introductory biology classroom and other racial/ethnic groups.

We can put the small achievement gap found in these biology classes into perspective by comparing our result with 1) achievement gaps based on other social identities for the same students, 2) achievement gaps in biology courses at other institutions, and 3) other studies of achievement gaps in college-level STEM courses. These provide a sense of how the magnitude of the gender gap compares with gaps that are already of concern in biology and whether or not biology is different from other STEM fields in terms of gender performance.

Social identities currently of concern in biology include first-generation status and racial/ethnic identity. We do not have data on first-generation status for our sample, but we do have racial and ethnic identity. Racial/ethnic achievement gaps are usually established by comparing a particular group's performance with white students. In our study, we found the difference in performance between males and females was similar in magnitude to that between white students and Latin@s, Native Americans, and Hawaiian and Pacific Islanders in these 23 classrooms. It was less than half of the achievement gap between white and black students and white domestic students and international students. The gender achievement gap was double that of the Asian and white achievement gap. These results reveal the gender achievement gap is of similar magnitudes to some gaps already of concern in biology, although it is smaller than others.

In contrast to our study, three studies in introductory biology classes found no significant achievement gaps between males and females (Willoughby and Metz, 2009; Creech and Sweeder, 2012; Lauer *et al*., 2013). However, these studies were only of one class each and thus substantially smaller in sample sizes than our study. In addition, only Creech and Sweeder (2012) controlled for student ability using college GPA as we did, and they found no gap in a 200-level biology class, but in a 400-level class females underperformed by 3.5% compared with males. Overall, our study is the largest study of introductory biology and the only study of introductory biology to demonstrate an achievement gap.

Compared with studies across STEM that also controlled for student prior academic performance when calculating a gender achievement gap, the achievement gap in biology we observed is slightly smaller in magnitude than in most other fields. These include studies in fields thought to be much less female friendly than biology, such as physics (7.5% lower; Miyake *et al*., 2010) and biochemistry (3.5-4.3%; Rauschenberger and Sweeder, 2010). The smaller achievement gap observed in our study implies that at least for our study, biology is different from other STEM fields in terms of female students' performance. Achievement gaps in performance are only one measure, though, and more measures need to be studied (and more institutions sampled) before any definitive conclusions can be drawn.

Explanations for achievement gaps between males and females in STEM are numerous, but, with our retrospective study design, we cannot distinguish among them. Instead, we will present two possible explanations, out of a myriad of possibilities (Hill *et al*., 2010), that seem plausible for our study setting. First, female students may enter introductory biology classes with a weaker biology background than males. Some evidence for this hypothesis comes from student scores on the Advanced Placement biology exam, on which males were found to consistently outperform females (Coley, 2001). Willoughby and Metz (2009) found that females underperformed on biology concept inventories given at the beginning of an introductory biology class. Additional evidence in support of a potential gap in preparedness for men and women can be found in other STEM fields. A gap in preparedness was found in a study of physics students in which females on average took fewer high school physics courses (Kost-Smith *et al*., 2010). More male high school students than female high school students have an interest in pursuing a STEM major (Ma, 2011), which could also lead to males taking more science courses in high school. Even if males and females took the same number of science classes in high school, females could still have a weaker background in biology if they did not receive the same opportunities to participate in STEM courses in K–12 that males did. There is evidence that males in K–12 classes are more likely to manipulate laboratory equipment and more likely to offer explanations in class, depriving females of opportunities to gain the skills that could be useful in college-level biology (Howe and Abedin, 2013). This difference in preparation, if present in our population, could potentially explain the achievement gap but needs to be further explored.

A second possible explanation for this achievement gap comes from the social psychology literature: the phenomenon of stereotype threat. Stereotype threat can be defined as fear that one's behaviors will confirm an existing stereotype of a group to which one belongs (Steele and Aronson, 1995). This phenomenon has been shown to reduce performance (Nguyen and Ryan, 2008) and is particularly strong in people who identify with the field they feel threatened in (e.g., identifying strongly with science; Inzlicht and Schmader, 2012). Interventions to alleviate stereotype threat have been shown to increase the performance of women in math-related fields (Spencer *et al*., 1999; Miyake *et al*., 2010). Currently, we do not have data on whether women are under stereotype threat in biology, although it is present across many other STEM fields (physics: Miyake *et al*., 2010; math: Spencer *et al*., 1999; computer science: Cheryan *et al*., 2009; engineering: Bell *et al*., 2003). Only one study has used a stereotype threat intervention in biology (Lauer *et al*., 2013), but this paper did not establish that there was an achievement gap between males and females before employing the intervention, making their negative result difficult to interpret; it is possible that the intervention could work in a classroom with a gender gap in

achievement. In addition, there are multiple types of stereotype threat (Inzlicht and Schmader, 2012), so the failure of one intervention that addresses one type of threat does not imply that other interventions will not work. Thus, it remains a possibility that women in biology are under stereotype threat and that this phenomenon could explain our results. Further work is needed to thoroughly explore this possibility.

In summary, we found a systematic achievement gap between males and females in our study, but, because our study design was retrospective, we had limited access to the measures necessary to distinguish between different explanations for the achievement gap we observed. Future prospective work could administer surveys that address differences in preparation and experience of stereotype threat to distinguish among these and other possibilities.

Instructor Gender May Impact the Achievement Gap

Evidence for an instructor gender effect on gender gaps in achievement at the college level is mixed. Some studies find that instructor gender does impact the achievement of females (Haley *et al*., 2007; Hoffman and Oreopoulos, 2009; Carrell *et al*., 2010), but other studies do not support this finding (Griffith, 2010; Price, 2010; Stout *et al*., 2011). Our study found some evidence for a small but significant impact of instructor gender, although there was some uncertainty about the importance of these terms (relative variable importance was moderate). Specifically, female students performed 0.1 of a SD better on exams when a course was taught exclusively by female instructors, which halved the achievement gap between males and females of the same ethnicity/race/nationality who entered the class with the same cumulative college GPA. This finding makes our study consistent and of a similar magnitude of effect with college-level STEM data that found that female students taught by female instructors in STEM courses outperformed female students taught by male instructors (Carrell *et al*., 2010). One limitation of our study is that we did not document whether teaching methods or exam format varied by instructor gender. Without this information, it is impossible to determine whether female instructors teach differently than male instructors and whether the instructor effect is due primarily to instructor gender. We do know anecdotally that the majority of exams across all 23 courses were short-answer format and that several of the instructors with the most student-centered classrooms were male.

Gender Gaps Exist in Whole-Class Participation

One of the novel aspects of this study is that we moved beyond simply quantifying academic achievement gaps to examining gaps in classroom participation in college-level STEM classrooms. Overall, we found that female and male students were equally likely to ask spontaneous questions in ~50% of the classes. When students were asked to offer volunteer responses, 69% of classrooms showed a pattern of male-biased participation; across these classes, males on average spoke 63% of the time, even though they comprised 40% of the overall class. Our study corroborates findings in elementary school science classrooms that show boys are eight times more likely to volunteer answers in class than girls (Sadker and Sadker, 1994). At the college level, studies of participation have found a range of patterns (more female than male: Howard and Henney, 1998; Howard *et al*., 2006; Fritschner, 2000; more male than female: Crombie *et al*., 2003; Tatum *et al*., 2013; no difference: Cornelius *et al*., 1990; Pearson and West, 1991; Brady and Eisler, 1999), but, to our knowledge, ours is the first observational study of college-level participation in a STEM classroom. In a study in STEM using self-reporting by students, women reported lower participation rates in biology, engineering, and chemistry courses (Crombie *et al*., 2003), and we have preliminary data showing a similar pattern in two introductory biology classrooms (unpublished data).

Class participation in our study took the form of interaction between two individuals: the instructor and the student. First, individual students decided whether or not to volunteer to answer an instructor's question, and then the instructor decided which volunteers to call on to speak. Either, or more likely both, individuals' behavior(s) could lead to the gender gap in participation observed in this study without anyone's conscious intent (Greenwald and Krieger, 2006).

Instructors enter their classroom with a set of perceptions about the class that may include, among many other things, what topics will interest students most, what students will already know about the subject, and who will participate the most. Some of these perceptions could include unconscious, and thus unexamined, biases about the roles of male and female students in the classroom and in science (Greenwald and Krieger, 2006). For example, if our previous experiences in science classrooms demonstrate that male students talk more and participate more actively than females (as shown in the K–12 literature: Holmes, 1992; Guzzetti and Williams, 1996; Howe and Abedin, 2013), then, as instructors, we might unconsciously expect the same pattern to occur in our classroom. Moreover, if we expect males to participate more, especially when offering answers (again seen in the K–12 literature; Altermatt *et al*., 1998; Burns and Myhill, 2004), then we might unconsciously facilitate this pattern by calling on males more. Thus, perpetuating gender inequality in the classroom can be a passive process that only requires us to remain unaware of our biased expectations (Greenwald and Krieger, 2006; Hill *et al*., 2010).

An illuminating example of this passive unconscious bias in a science classroom comes from a study at the elementary school level in which researchers worked with science instructors to equalize student participation. Instructors involved in this process found it difficult, and one instructor reported that he felt he was devoting 90% of class time to females, when really it was just equal time (Whyte, 1986). It was his unconscious bias that females would not participate at equal rates that influenced his perception of the classroom dynamics. A more recent study demonstrating that this unconscious bias against women persists in STEM found that faculty members of all genders were more likely to hire a male undergraduate lab assistant than a female, pay the male lab assistant more than the female, and offer a greater level of mentoring, even when the candidates had identical qualifications (Moss-Racusin *et al*., 2012).

The second factor that could contribute to the gender bias in participation in volunteer-based classroom interactions is a student's decision to volunteer. In the K–12 literature, there is a consistent pattern wherein females speak less than males

in traditionally male-dominated fields such as science. There is also extensive evidence at the K–12 level that girls are less confident than boys in their knowledge in science fields even after controlling for their actual performance (Meece *et al.*, 2006; Micari *et al.*, 2007; Sikora and Pokropek, 2012) and, thus, may not feel confident enough to provide an answer in front of a large group. Girls also seem to be much more concerned about how their instructors view them, and the fear of creating a negative perception could hold them back from participating (Pomerantz *et al.*, 2002). At the college level, this difference in confidence between males and females has been demonstrated in several STEM disciplines, including engineering (Cech *et al.*, 2011) and physics (Lindstrom and Sharma, 2011).

When participation in biology classes is skewed toward males, females are systematically missing out on valuable practice that may lead to benefits such as achievement and/or retention in STEM. Although these are often the more common benchmarks for success in education research, speaking in class can also strengthen a student's relationship with the field of biology and improve his or her sense of belonging, which could indirectly impact retention. For example, speaking and earning praise, or hearing people with a similar social identity (e.g., same gender or race) as you earning praise from an authority figure has been shown to increase a student's sense of belonging in a field (Carlone and Johnson, 2007; Ong *et al.*, 2011). By not being called on and not receiving the validation of an authority figure (e.g., the instructor; Sinnes and Loken 2012), females may develop a lower sense of belonging as a person who can contribute to the biology community. This incongruence between how a female views herself as a person capable of being a competent biologist and how she thinks others view her could lead to stereotype threat or imposter syndrome, the conviction that despite her accomplishments she is still not good enough for the field. Both of these phenomena are known to decrease student performance and contribute to attrition from STEM fields (Massey and Fischer, 2005; Freeman *et al.*, 2007b; Collett *et al.*, 2013).

In addition, in biology and STEM fields in general, practitioners must be comfortable offering their ideas in group settings such as meetings, conferences, and day-to-day interactions with collaborative teams. Science classrooms are an opportunity for students to practice these skills in a low-stakes environment. At first it may seem unkind to call on students who are hesitant to participate, but research has shown that students who participate in class, even if they are forced to participate initially through cold-calling, become more comfortable talking in class and even begin to volunteer on their own (Dallimore *et al.*, 2010, 2013). This increased confidence in participating could transfer to higher-stakes environments such as lab meetings and scientific conferences. Thus, the limited participation of females in introductory biology classrooms is denying females the chance to practice science discourse skills to the same degree as males and preventing them from gaining the confidence to participate in more high-stakes environments. Furthermore, classrooms in which males dominate discussions could indicate to future male scientists that underparticipation by females in biology is standard. For all these reasons, unequal class participation may have greater and more enduring consequences for equity that are difficult to measure.

Disparities in Whole-Class Discussions Can Be Ameliorated Using Random Call

It seems that factors at both the student and instructor level could lead to disparities in who participates in the biology classroom. Fortunately, our results also indicate that there is at least one simple solution to the problem: using random call to structure class participation. By choosing to employ this interaction method, instructors will call on males and females in proportion to their representation in class and can prevent gender disparity in who participates. Random call differs from volunteer-based student participation, because it requires the instructor to call on people based on a list created before he or she enters the classroom. This list of randomized names not only prevents any instructor bias from influencing who is called on but also does not allow students to opt out of participation because they are uncomfortable. Random call may sound intimidating to students, but instructors in this study alleviated this anxiety somewhat by having students discuss their answers in small groups (e.g., think–pair–share) before anyone was called on to report to the whole class. Random call is additionally useful, because it spreads participation equally over the whole class and prevents a few students from monopolizing an instructor's time. In this study, instructors made a randomized class list in advance using Microsoft Excel, but other instructors have used a deck of cards with students names on them that they shuffled and drew from (Tanner, 2013), and there are even apps for modern handheld devices designed for this task, such as Names in a Hat for iPhone.

Replication with Other Student Populations Is Necessary before Drawing Conclusions across Biology

In this study, we focused on two measures of gender equity in biology across 23 classes with almost 5000 students. Although our sample size is large, the entire sample is taken from only one institution that has its own unique identity, culture, and student demographics. Gender inequity in biology is a complex factor that is influenced by the experiences brought to the classroom by both the students and the instructors. Therefore, it is important for researchers not to make assumptions about the dynamics of social identities in biology classrooms based only on data from this paper, which represent only one institution. Rather, both researchers and instructors need to document the gender patterns in their own classes and institutions to determine the pervasiveness of gender gaps. In addition, unconscious bias and learned roles in the classroom are experiences that could influence a range of student outcomes, including self-efficacy, interest in science, and course-related anxiety. In this study, we only investigated two of the myriad of measures that could be used to elucidate areas of gender inequities in biology; there is the need for more work to be done to identify patterns in the experience of females in biology across a range of outcome variables and a range of institution types.

CONCLUSION

Although biology has been successful at closing gender gaps in attracting and retaining undergraduate and graduate students, in this study, we document more subtle gaps

that persist. We found that both academic achievement and participation in class reveal evidence of systematic gender differences in introductory biology at a R1 institution, which suggest that there may be many unexplored aspects of science identity development that remain to be addressed before we can purport gender equity in biology.

As the undergraduate student body continues to diversify at colleges universities, it is becoming increasingly important that instructors not only have deep content expertise and use evidence-based teaching practices, but also that they are aware of the challenges facing students of different social identities in the biology classroom. Many of these barriers have already been identified by the social sciences; these researchers have also developed many successful interventions to help students cope with lower confidence (Aronson *et al*., 2002), stereotype threat (Cohen *et al*., 2006; Miyake *et al*., 2011), and other barriers in the classroom faced by specific groups. As we work toward improving undergraduate biology education for all students, recognizing and challenging our own biases is an essential first step toward making undergraduate biology more equitable. The remaining challenge for all of us is to act on our awareness by modifying our teaching to maximize the learning environment for the ever increasing diversity of students in our classrooms.

ACKNOWLEDGMENTS

Support for this study was provided by NSF TUES 1118890. We thank Chessa Goss, John Parks, Ben Wiggins, and Leslie Zeman for helping us gather all the records for this study; Mercedes Converse, Carl Longton, and Michael Mullen for help coding videos; and Alison Crowe, Scott Freeman, Dan Grunspan, Margaret Blankenbiller, Kate Boersma, Chris Lenn, Steve Kroiss, and the Biology Education Research Group for their comments on earlier drafts of the manuscript. This research was done under approved IRB #38945, University of Washington.

REFERENCES

Akaike H (1973). Information theory as an extension of the maximum likelihood principle. In: Second International Symposium on Information Theory, ed. BN Petrov and F Csaki, Budapest: Akademiai Kiado, 267–281.

Akaike H (1974). A new look at the statistical model identification. IEEE Trans Autom Control *19*, 716–723.

Altermatt ER, Jovanovic J, Perry M (1998). Bias or responsivity? Sex and achievement-level effects on teachers' classroom questioning practices. J Educ Psychol *90*, 516–527.

Amelink C (2009). Literature Overview: Gender Differences in Science Achievement, SWE-AWE CASEE Overviews, University Park, PA: Assessing Women and Men in Engineering Project.

Anderson DR (2008). Model Based Inference in the Life Sciences: A Primer on Evidence, New York: Springer.

Anderson DR, Burnham KP, Thompson WL (2000). Null hypothesis testing: problems, prevalence, and an alternative. J Wildl Manag *64*, 912–923.

Anderson ML (2005). Thinking about women a quarter century's view. Gender Soc *19*, 437–455.

Antecol H, Eren O, Ozbeklik (2012). The Effect of Teacher Gender on Student Achievement in Primary School: Evidence from a Randomized Experiment, Discussion Paper no. 6453, Bonn: Institute for the Study of Labor.

Aronson J, Fried CB, Good C (2002). Reducing the effects of stereotype threat on African American college students by shaping theories of intelligence. J Exp Soc Psychol *38*, 113–125.

Barton K (2013). MuMIn: Multi-model Inference, R package version 1.9.5. http://CRAN.R-project.org/package=MuMIn (accessed 27 October 2013).

Bates D, Maechler M, Bolker B, Walker S (2013). lme4: Linear Mixed-effects Models Using Eigen and S4. R package version 1.0-4. http://cran.r-project.org/web/packages/lme4/index.html (accessed 27 October 2013).

Beasley MA, Fischer MJ (2012). Why they leave: the impact of stereotype threat on the attrition of women and minorities from science, math and engineering majors. Soc Psychol Educ *15*, 427–448.

Beede D, Julian T, Langdon D, McKittrick G, Khan B, Doms M (2011). Women in STEM: A Gender Gap to Innovation, Economics and Statistics Administration Issue Brief 04–11, Washington, DC: U.S. Department of Commerce.

Beichner RJ, Saul JM, Abbott DS, Morse JJ, Deardorff DJ, Allain RJ, Bonham SW, Dancy MH, Risley JS (2007). The student-centered activities for large enrollment undergraduate programs (SCALE-UP) project. In: Research-based Reform of University Physics, vol. 1, ed. EF Redish and PJ Cooney, College Park, MD. http://www.percentral.com/PER/per_reviews/media/volume1/SCALE-UP-2007.pdf (accessed 5 February 2014).

Bell AE, Spencer SJ, Iserman E, Logel CER (2003). Stereotype threat and women's performance in engineering. J Eng Educ *92*, 307–312.

Bell M, Gafni N (2000). Can item format (multiple choice vs. open-ended) account for gender differences in mathematics achievement? Sex Roles *42*, 1–21.

Bolker BM, Brooks ME, Clark CJ, Geange SW (2009). Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol Evol *24*, 127–135.

Brady KL, Eisler RM (1999). Sex and gender in the college classroom: a quantitative analysis of faculty-student interactions and perceptions. J Educ Psychol *91*, 127–145.

Brownell SE, Kloser MJ, Fukami T, Shavelson RJ (2013). Context matters: volunteer bias, small sample size, and the value of comparison groups in the assessment of research-based undergraduate introductory biology lab courses. J Microbiol Biol Educ *14*, 176–182.

Burke RJ, Mattis MC (eds.) (2007). Women and Minorities in Science, Technology, Engineering, and Mathematics: Upping the Numbers, Cheltenham, UK: Edward Elgar.

Burnham KP, Anderson DR (2004). Multimodel inference: understanding AIC and BIC in model selection. Sociol Method Res *33*, 261–304.

Burnham KP, Anderson DR, Huyvaert KP (2011). AIC model selection and multimodal inference in behavioral ecology: some background, observations, and comparisons. Behav Ecol Sociobiol *65*, 23–35.

Burns C, Myhill D (2004). Interactive or inactive? A consideration of the nature of interaction in whole class teaching. Cambridge J Educ *34*, 35–49.

Carlone HB, Johnson A (2007). Understanding the science experiences of women of color: science identity as an analytical lens. J Res Sci Teach *44*, 1187–1218.

Carrell SE, Page ME, West JW (2010). Sex and science: how professor gender perpetuates the gender gap. Q J Econ *125*, 1101–1144.

Cech E, Rubineau B, Silbey S, Seron C (2011). Professional role confidence and gender persistence in engineering. Am Soc Rev *76*, 641–666.

Ceci SJ, Williams WM (2010). Gender differences in math-intensive fields. Curr Dir Psychol Sci *19*, 275–279.

Ceci SJ, Williams WM (2011). Understanding current causes of women's underrepresentation in science. Proc Natl Acad Sci USA *108*, 3157–3162.

Cheryan S, Plaut VC, Davies P, Steele CM (2009). Ambient belonging: how stereotypical environments impact gender participation in computer science. J Person Soc Psychol *97*, 1045–1060.

Clark Blickenstaff J (2005). Women and science careers: leaky pipeline or gender filter? Gender Educ *17*, 369–389.

Cohen GL, Garcia J, Apfel N, Master A (2006). Reducing the racial achievement gap: a social-psychological intervention. Science *313*, 1307–1310.

Coley RJ (2001). Differences in the gender gap: comparisons across racial/ethnic groups in education and work. Educational Testing Service. www.ets.org/research/pic/gender.pdf (accessed 15 February 2014).

Collett JL, Avelis J, Lizardo O (2013). Family-Friendliness, Fraudulence, and Gendered Academic Career Ambitions. Paper presented at the American Sociological Association Annual Meeting, held 11 August 2013, in New York City.

Cornelius RR, Gray JM, Constantinople AP (1990). Student-faculty interaction in the college classroom. J Res Develop Educ *23*, 189–197.

Creech LR, Sweeder RD (2012). Analysis of student performance in large-enrollment life science courses. CBE Life Sci Educ *11*, 386–391.

Crombie G, Pyke SW, Silverthorn N, Jones A, Piccinin S (2003). Students' perceptions of their classroom participation and instructor as a function of gender and context. J High Educ *74*, 51–76.

Dallimore EJ, Hertenstein JH, Platt MB (2010). Class participation in accounting courses: factors that affect student comfort and learning. Issues Account Educ *25*, 613–629.

Dallimore EJ, Hertenstein JH, Platt MB (2013). Impact of cold-calling on student voluntary participation. J Manage Educ *37*, 305–341.

Demby G (2013). Latin@ offers a gender neutral choice; But how to pronounce it? www.npr.org/blogs/thetwo-way/2013/01/07/168818064/latin-offers-a-gender-neutral-choice-but-how-to-pronounce-it (accessed 25 May 2014).

Dimitrov DD (1999). Gender differences in science achievement: differential effect of ability, response format and strands of learning outcomes. Sch Sci Math *99*, 445–450.

Ecklund EH, Lincoln AE, Tansey C (2012). Gender segregation in elite academic science. Gender Soc *26*, 693–717.

Fassinger PA (2000). How classes influence students' participation in college classrooms. J Class Inter *35*, 38–47.

Ferreira M (2003). Gender issues related to graduate student attrition in two science departments. Int J Sci Educ *25*, 969–989.

Freeman S, Haak D, Wenderoth MP (2011). Increased course structure improves performance in introductory biology. CBE Life Sci Educ *10*, 175–186.

Freeman S, O'Conner E, Parks JW, Cunningham M, Hurley D, Haak D, Dirks C, Wenderoth MP (2007a). Prescribed active learning increases performance in introductory biology. CBE Life Sci Educ *6*, 132–139.

Freeman TM, Anderman LH, Jensen JM (2007b). Sense of belonging in college freshmen at the classroom and campus levels. J Exp Educ *75*, 203–220.

Fritschner LM (2000). Inside the undergraduate classroom: faculty and students differ on the meaning of student participation. J High Educ *71*, 342–362.

Garamszegi LZ (2011). Information-theoretic approaches to statistical analysis in behavioral ecology: an introduction. Behav Ecol Sociobiol *65*, 1–11.

Greenwald AG, Krieger LH (2006). Implicit bias: scientific foundations. Calif Law Rev *94*, 945–967.

Griffith AL (2010). Persistence of women and minorities in STEM field majors: is it the school that matters? Econ Educ Rev *29*, 911–922.

Guzzetti BJ, Williams WO (1996). Gender, text, and discussion: examining intellectual safety in the science classroom. J Res Sci Teach *33*, 5–20.

Haak DC, HilleRisLambers J, Pitre E, Freeman S (2011). Increased structure and active learning reduce the achievement gap in introductory biology. Science *332*, 1213–1216.

Haley MR, Johnson MF, Kuennen EW (2007). Student and professor gender effects in introductory business statistics. J Stat Educ *15*, 1–19.

Handelsman J (2005). More women in science. Science *309*, 1190–1191.

Hegyi G, Garamszegi LZ (2011). Using information theory as a substitute for stepwise regression in ecology and behavior. Behav Ecol Sociobiol *65*, 69–76.

Hill C, Corbett C, St. Rose A (2010). Why So Few? Women in Science, Technology, Engineering, and Mathematics, Washington, DC: American Association of University Women.

Hoffmann F, Oreopoulos P (2009). A professor like me: the influence of instructor gender on college achievement. J Hum Resour *44*, 479–494.

Holmes J (1992). Women's talk in public contexts. Discourse Soc *3*, 131–150.

Howard JR, Henney AL (1998). Student participation and instructor gender in the mixed-age college classroom. J High Educ *69*, 384–405.

Howard JR, Zoeller A, Pratt Y (2006). Students' race and participation in sociology classroom discussion: a preliminary investigation. J Scholarship Teach Learn *5*, 14–38.

Howe C, Abedin M (2013). Classroom dialogue: a systematic review across four decades of research. Cambridge J Educ *43*, 325–356.

Inzlicht M, Schmader T (2012). Stereotype Threat: Theory, Process, and Application, New York: Oxford University Press.

Isbell LA, Young TP, Harcourt AH (2012). Stag parties linger: continued gender bias in a female-rich scientific discipline. PLoS One *7*, e49682.

Johnson AC (2007). Unintended consequences: how science professors discourage women of color. Sci Educ *91*, 805–821.

Johnson JB, Omland KS (2004). Model selection in ecology and evolution. Trends Ecol Evol *19*, 101–108.

Jovanovic J, King SS (1998). Boys and girls in the performance-based science classroom: who's doing the performing? Am Educ Res J *35*, 477–496.

Kane TJ, Staiger DO (2012). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains, Policy and Practice Brief, MET Project, Seattle, WA: Bill & Melinda Gates Foundation.

Kost-Smith LE, Pollock SJ, Finkelstein ND (2010). Gender disparities in second-semester college physics: the incremental effects of a "smog of bias." Phys Educ Res *6*, 1–17.

Kreft IGG, de Leeuw J (2002). Introducing Multilevel Modeling, Thousand Oaks, CA: Sage.

Lane KA, Goh JX, Driver-Linn E (2012). Implicit science stereotypes mediate the relationship between gender and academic participation. Sex Roles *66*, 220–234.

Lariviere V, Ni C, Gingras Y, Cronin B, Sugimoto CR (2013). Bibliometrics: global gender disparities in science. Nature *504*, 211–213.

Lauer S, Momsen J, Offerdahl E, Kryjevskaia WC, Montplaisir L (2013). Stereotyped: investigating gender in introductory science courses. CBE Life Sci Educ *12*, 30–38.

Lindstrøm C, Sharma MD (2011). Self-efficacy of first year university physics students: do gender and prior formal instruction in physics matter? Int J Innov Sci Math Educ *19*, 1–19.

Lipsey MW, Puzio K, Yun C, Hebert MA, Steinka-Fry K, Cole MW, Roberts M, Anthony KS, Busick MD (2012). Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms (NCSER 2013–3000), Washington, DC: U.S. Department of Education.

Lorenzo M, Crouch CH, Mazur E (2006). Reducing the gender gap in the physics classroom. Am J Phys *74*, 118–122.

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

Ma Y (2011). Gender differences in the paths leading to a STEM baccalaureate. Soc Sci Q *92*, 1169–1190.

Massey DS, Fischer MJ (2005). Stereotype threat and academic performance: new findings from a racially diverse sample of college freshmen. Du Bois Rev *2*, 45–67.

McQuarrie ADR, Tsai C-L (1998). Regression and Time Series Model Selection, London: World Scientific.

Meece JL, Glienke BB, Burg S (2006). Gender and motivation. J School Psychol *44*, 351–373.

Micari M, Pazos P, Hartmann MJ (2007). A matter of confidence: gender differences in attitudes toward engaging in lab and course work in undergraduate engineering. J Women Minor Sci Eng *13*, 279–293.

Miyake A, Kost-Smith LE, Finkelstein ND, Pollock SJ, Cohen GL, Ito TA (2010). Reducing the gender achievement gap in college science: a classroom study of values affirmation. Science *330*, 1234–1237.

Moss-Racusin CA, Dovidio JF, Brescoll VL, Graham MJ, Handelsman J (2012). Science faculty's subtle gender biases favor male students. Proc Natl Acad Sci USA *109*, 16474–16479.

Mulvey PJ, Nicholson S (2011). Focus On—Physics Enrollments: Results from the 2008 Survey of Enrollments and Degrees, College Park, MD: Statistical Research Center of the American Institute of Physics.

Mulvey PJ, Nicholson S (2012). Focus On—First-Year Physics Graduate Students: Characteristics and Background. Data from the 2007–2008 and 2009–2010 Graduate Student Surveys, College Park, MD: Statistical Research Center of the American Institute of Physics.

National Research Council (2007). Beyond Bias and Barriers: Fulfilling the Potential of Women in Academic Science and Engineering, Washington, DC: National Academies Press.

National Science Foundation (2011). Women, Minorities, and Persons with Disabilities in Science and Engineering: 2011 (NSF 11-309), Arlington, VA: National Science Foundation. www.nsf.gov/statistics/wmpd (accessed 27 October 2013).

Nguyen H-H D, Ryan AM (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. J Appl Psychol *93*, 1314–1334.

Nosek BA, Banaji MR, Greenwald AG (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. Group Dynam *6*, 101–115.

Nunn CE (1996). Discussion in the college classroom: triangulating observational and survey results. J High Educ *67*, 243–266.

Ong M, Wright C, Espinosa LL, Orfield G (2011). Inside the double bind: a synthesis of empirical research on undergraduate and graduate women of color in science, technology, engineering and mathematics. Harvard Educ Rev *81*, 172–208.

Paterson L, Goldstein H (1991). New statistical methods for analysing social structures: an introduction to multilevel models. Brit Educ Res J *17*, 387–393.

Pearson JC, West R (1991). An initial investigation of the effects of gender on student questions in the classroom: developing a descriptive base. Commun Educ *40*, 22–32.

Pomerantz EM, Altermatt ER, Saxon JL (2002). Making the grade but feeling distressed: gender differences in academic performance and internal distress. J Educ Psychol *94*, 396–404.

Price J (2010). The Effect of Instructor Race and Gender on Student Persistence in STEM Fields (Working Paper 121), Ithaca, NY: Cornell University, School of Industrial and Labor Relations. http://digitalcommons.ilr.cornell.edu/workingpapers/121 (accessed 10 October 2013).

Raudenbush SW, Bryk AS (2002). Hierarchical Linear Models: Applications and Data Analysis Methods, 2nd ed., Thousand Oaks, CA: Sage.

Rauschenberger MM, Sweeder RD (2010). Gender performance differences in biochemistry. Biochem Mol Biol Educ *38*, 380–384.

R Core Team (2012). R: A Language and Environment for Statistical Computing, Vienna, Austria: R Foundation for Statistical Computing. www.R-project.org (accessed 27 October 2013).

Riegle-Crumb C, King B (2010). Questioning a white male advantage in STEM: examining disparities in college major by gender race/ethnicity. Educ Res *39*, 656–664.

Riegle-Crumb C, King B, Grodsky E, Muller C (2012). The more things change, the more they stay the same? Prior achievement fails to explain gender inequality in entry into STEM college majors over time. Am Educ Res J *49*, 1048–1073.

Riegle-Crumb C, Moore C, Ramos-Wada A (2011). Who wants to have a career in science or math? Exploring adolescents' future aspirations by gender and race/ethnicity. Sci Educ *95*, 458–476.

Rosser SV (2012). Breaking into the Lab: Engineering Progress for Women in Science, New York: New York University Press.

Sadker D, Sadker M (1994). Failing at Fairness: How Our Schools Cheat Girls, Toronto, ON: Simon & Schuster.

Schmader T, Johns M, Barquissau M (2004). The cost of accepting gender differences: the role of stereotype endorsement in women's experience in the math domain. Sex Roles *50*, 835–850.

Sikora J, Pokropek A (2012). Gender segregation of adolescent science career plans in 50 countries. Sci Educ *96*, 234–264.

Sinnes AT, Loken M (2012). Gendered education in a gendered world: looking beyond cosmetic solutions to the gender gap in science. Cult Stud Sci Educ *2012*, 1–22.

Smith MK, Jones FHM, Gilbert SL, Wieman CE (2013). The Classroom Observation Protocol for Undergraduate STEM (COPUS): a new instrument to characterize university STEM classroom practices. CBE Life Sci Educ *12*, 618–627.

Spencer JS, Steele CM, Quinn DM (1999). Stereotype threat and women's math performance. J Exp Soc Psychol *35*, 4–28.

Steele CM, Aronson J (1995). Stereotype threat and the intellectual test performance of African Americans. J Person Soc Psych *69*, 797–811.

Stout JG, Dasgupta N, Hunsinger M, McManus MA (2011). STEMing the tide: using in group experts to inoculate women's self-concept in science, technology, engineering, and mathematics (STEM). J Pers Soc Psychol *100*, 255–270.

Strenta AC (1994). Choosing and leaving science in highly selective institutions. Res High Educ *35*, 513–547.

Symonds MRE, Moussalli A (2011). A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. Behav Ecol Sociobiol *65*, 13–21.

Tanner KD (2013). Structure matters: twenty-one teaching strategies to promote student engagement and cultivate classroom equity. CBE Life Sci Educ *12*, 1–10.

Tatum HE, Schwartz BM, Schimmoeller PA, Perry N (2013). Classroom participation and student-faculty interactions: does gender matter? J High Educ *84*, 745–768.

Theobald R, Freeman S (2014). Is it the intervention or the students? Using linear regression to control for student characteristics in undergraduate STEM education research. CBE Life Sci Educ *13*, 41–48.

Tsui L (2002). Fostering critical thinking through effective pedagogy: evidence from four institutional case studies. J High Educ *73*, 740–763.

Whyte J (1986). Girls into Science and Technology, London: Routledge & Kegan Paul.

Wickware P (1997). Along the leaky pipeline. Nature *390*, 202–203.

Willoughby SD, Metz A (2009). Exploring gender differences with different gain calculations in astronomy. Am J Phys *77*, 651–657.

Xie Y, Shauman KA (2003). Women in Science: Career Processes and Outcomes, Cambridge, MA: Harvard University Press.

Young DM, Rudman LA, Buettner HM, McLean MC (2013). The influence of female role models on women's implicit science cognitions. Psychol Women Q *37*, 283–292.

Zuur AF, Ieno EN, Walker NJ, Saveliev AA, Smith GM (2009). Mixed Effect Models and Extensions in Ecology in R, New York: Springer.

HIGHLIGHT:

Although females outnumber males in biology, this study of 23 different introductory biology classrooms reveals systematic gender disparities in student performance on exams and student participation when instructors ask students to volunteer answers to instructor-posed questions.