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

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.

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 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.
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 150 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@2, 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, and 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 among 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
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, it's 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; Akaikes, 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 (Akaikes, 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 (Akaikes, 1973; Anderson, 2008). These AICc values are then used to rank the models. From these AICc values, AICc differences (Δi) and Akaikes weights (ωi) 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 Δi > 10 are considered poor predictors compared with the best model and thus are not included in our analyses (Burnham and Anderson, 2004). Akaikes weights (ωi) 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. Akaikes 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 Akaikes 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:

\[
\text{Z.Score} = \text{Cum.GPA} + \text{Stu.Gender} + \text{Ethn} + \text{Stu.Gender} \times \text{Ethn} + \text{Inst.Gender} + (\text{1/Stu.ID}) + (\text{Cum.GPA} \times \text{Class})
\]

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 (Δi < 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 ω = 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 variable intercepts (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 (β = 0.21 ± 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 (ω1 = 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@ (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

<table>
<thead>
<tr>
<th>Rank</th>
<th>Modela</th>
<th>AICc</th>
<th>Δi</th>
<th>ω1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cum.GPA + Ethn + Stu.Gender</td>
<td>18019.9</td>
<td>0</td>
<td>0.41</td>
</tr>
<tr>
<td>2</td>
<td>Cum.GPA + Ethn + Stu.Gender + Inst.Gender + Stu.Gender*Inst.Gender</td>
<td>18020.6</td>
<td>0.63</td>
<td>0.30</td>
</tr>
<tr>
<td>3</td>
<td>Cum.GPA + Ethn + Stu.Gender + Inst.Gender</td>
<td>18022.5</td>
<td>2.58</td>
<td>0.11</td>
</tr>
<tr>
<td>4</td>
<td>Cum.GPA + Ethn + Stu.Gender + Ethn</td>
<td>18023.1</td>
<td>3.21</td>
<td>0.08</td>
</tr>
<tr>
<td>5</td>
<td>Cum.GPA + Ethn + Stu.Gender + Stu.Gender<em>Ethn + Inst.Gender + Stu.Gender</em>Inst.Gender</td>
<td>18023.5</td>
<td>1.95</td>
<td>0.07</td>
</tr>
<tr>
<td>6</td>
<td>Cum.GPA + Ethn + Stu.Gender + Inst.Gender</td>
<td>18025.7</td>
<td>5.79</td>
<td>0.02</td>
</tr>
</tbody>
</table>

aRelative ranking (from most support to least) of six best models for predicting student exam performance using AICc model selection. Only models that are informative (Δi < 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.

bCum.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.
Instructor gender and the interaction between instructor gender and student gender were present in the second-best model (Table 1) and have relative variable importance 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 independent observers, one male and one female, who recorded 1) the conversations through our whole-class video recordings. 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 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.
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 $61.0 \pm 0.04\%$ of students in the class were female and $60.0 \pm 11.8\%$ of the participants in random call were female (Figure 3). This pattern was consistent across the four classrooms in which random call was used.

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@es. Replication of this difference and more detailed studies will be necessary to parse out the significance of the difference between the experience of Latin@es 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@es, 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 Eiesler, 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, 2002; 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 confident 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 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; Merced 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


Amelink C (2009). Literature Overview: Gender Differences in Science Achievement, SWE-AWE CASEE Overviews, University Park


Antecol H, Eren O, Ozbeklik (2012). The Effect of Teacher Gender on Student Achievement in Primary School: Evidence from a Random-

Gender Gaps in Introductory Biology


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.


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.