Smaller Classes Promote Equitable Student Participation in STEM

CISSY J. BALLEN, STEPFANIE M. AGUILLON, AZZA AWWAD, ANNE E. BJUNE, DANIEL CHALLOU, ABBY GRACE DRAKE, MICHELLE DRIESSEN, AZIZA ELLOZY, VIVIAN E. FERRY, EMMA E. GOLDBERG, WILLIAM HARCOMBE, STEVE JENSEN, CHRISTIAN JØRGENSEN, ZOE KOTH, SUZANNE MCGAUGH, CAROLINE MITRY, BRYAN MOSHER, HODA MOSTAFA, RENEE H. PETIPAS, PAULA A. G. SONERAL, SHANA WATTERS, DEENA WASSENBERG, STACEY L. WEISS, AZARIAH YONAS, KELLY R. ZAMUDIO, AND SEHOYA COTNER

As science, technology, engineering, and mathematics (STEM) classrooms in higher education transition from lecturing to active learning, the frequency of student interactions in class increases. Previous research documents a gender bias in participation, with women participating less than would be expected on the basis of their numeric proportions. In the present study, we asked which attributes of the learning environment contribute to decreased female participation: the abundance of in-class interactions, the diversity of interactions, the proportion of women in class, the instructor's gender, the class size, and whether the course targeted lower division (first and second year) or upper division (third or fourth year) students. We calculated likelihood ratios of female participation from over 5300 student–instructor interactions observed across multiple institutions. We falsified several alternative hypotheses and demonstrate that increasing class size has the largest negative effect. We also found that when the instructors used a diverse range of teaching strategies, the women were more likely to participate after small-group discussions.

Keywords: STEM equity, gender, in-class interactions, class size, active learning

Active learning can be distinguished from traditional lecturing through its emphasis on diverse types of engagement strategies, including structured student–instructor interactions during activities and guided inquiry (Smith et al. 2009, Haak et al. 2011). Substantial evidence supports interactive classes as a more effective form of instruction than traditional lecture (Freeman et al. 2014), particularly for at-risk students (Lorenzo et al. 2006, Beichner et al. 2007, Haak et al. 2011, Ballen et al. 2017). However, the most effective and equitable types of interactions that support all students in their learning are a subject of current debate. This question is particularly critical in gateway courses that are required for all students before they can pursue more specialized coursework. Across the science, technology, engineering, and mathematics (STEM) disciplines, students struggle in gateway courses, and failure rates are high (Freeman et al. 2011, National Academies of Sciences and Medicine 2016). Therefore, it is critical that gateway courses be systematically assessed to identify which elements within the classrooms leads to gaps in participation and which elements provide structure when needed.

Previous research demonstrates a pervasive gender gap in participation in undergraduate STEM courses (Eddy et al. 2014), a trend that persists beyond undergraduate lecture halls. In fact, it has been shown that women audience members ask fewer questions than men after academic seminar and conference talks (Pritchard et al. 2014, Carter et al. 2017, Hinsley et al. 2017). These patterns may contribute to a general tendency to undervalue the contributions of women and lead to documented phenomena such as proportionately fewer women awarded prestigious fellowships (Wold and Wenneras 2010) and grants (Ledin et al. 2007), fewer female first (primary investigators; O'Dorchai et al. 2009) and last authors (research supervisors; Holman et al. 2018), fewer women invited as speakers at symposia (Isbell et al. 2012), and fewer women occupying high-status positions in STEM (O'Dorchai et al. 2009, Beede et al. 2011). Therefore, factors that contribute to unequal participation should be identified and proper interventions should be designed early in STEM education.

Variability in female participation across classrooms indicates the presence of underlying, course-specific factors that create environments more or less encouraging to the input of women. We selected six course elements from the literature that may affect female participation and used deductive methods to understand each element's relative impact on equitable participation from our sample of observations (table 1).

BioScience XX: 1–12. © The Author(s) 2019. Published by Oxford University Press on behalf of the American Institute of Biological Sciences. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com.
doi:10.1093/biosci/biz069

Table 1. Alternative hypotheses that may explain, in isolation or in combination, equitable in-class participation in STEM courses.

Predictor	Reasoning: Students may be more comfortable speaking in class…
Abundance of student–instructor interactions per class period	…if participation is normalized through many different instances of student–instructor interactions throughout class (Kuh and Hu 2001, Komarraju et al. 2010).
Diversity of interactions	…if the instructor uses a wide range of teaching strategies, generally involving peer discussions, (e.g., small-group discussions, classroom response systems, think–pair–share) intended to encourage equitable participation (Premo and Cavagnetto 2018).
Instructor gender	…if the gender of the instructor matches their own (Crombie et al. 2003, Cotner et al. 2011).
Proportion of women in the class	…if genders are represented in relatively equitable proportions, so that the underrepresented gender does not feel isolated in the larger social setting (Dahlerup 1988).
Class size	…if they are in a classroom with fewer students (Kokkelenberg et al. 2008, Schanzenbach 2014, Ballen et al. 2018a).
Lower division or upper division	…if they are in an upper division course, having cleared the hurdle of the introductory, weed-out courses (Brewer and Smith 2011). Alternatively, students warmed to instructional methods over time, including in-class activities.

We examined how the abundance of interactions, the diversity of interactions, the instructor's gender, the proportion of women in the class, the class size, and the class division affect three specific types of student participation: voluntary responses, when an instructor poses a question and an individual raises their hand to answer without conferring with their peers; group responses, when an instructor poses a question and students have the opportunity to talk to their peers before answering; and total responses, or all student–instructor interactions observed across a class period. A summary of our reasoning for several hypotheses (predictors) for female participation is provided in table 1. We addressed the following research question as it applies across multiple universities: What leads to gendered participation in science classrooms in higher education? We developed a number of alternative hypotheses that might predict why, in some environments, we observe individuals of one gender speaking more than another (table 1).

Participant and institutional information

We collected student behavioral data from 44 courses across the United States. As part of the creation of this larger collaborative research group, we solicited participation through an existing professional network from instructors who teach majors, nonmajors, or both at a range of institutions. The volunteers represented Bethel University, in Saint Paul, Minnesota; Cornell University, in Ithaca, New York; the University of Minnesota (UNM), in Saint Paul; the University of Puget Sound, in Tacoma, Washington; the American University in Cairo, Egypt; and the University of Bergen, in Norway (table 2). The participating institutions were a convenience sample chosen from a range of institutional types (public and private, large and small) and settings (college towns to large metropolitan areas). During the 2-year study period, approximately 5200 students enrolled in the sampled courses, and observers categorized over 5300 interactions between the instructors and the students We included courses from across STEM fields, including biology, physics, computer science, and chemistry (details in the raw data file in the supplemental material). Demographic information collected by university registrars revealed that, on average, 53.8% of the students in these classes identified as female, but this number ranged from 20.4% to 79.6%, depending on the specific class. All aspects of research were reviewed and approved by each schools' respective institutional review board (Bethel IRB 180,518; Cornell IRB 1,410,005,010; UNM IRB 00000800; University of Puget Sound IRB 1617–006; American University in Cairo 2016–2017-0012; University of Bergen NSD 46,727).

Measuring in-class participation

We conducted training sessions of approximately 1 hour for the observers to characterize classroom participation as broad types of interactions that occur over a class period, which were further characterized as either *voluntary responses* or *group responses*. For each type of interaction that took place during a class period, an observer recorded the gender of the student participant (0, female; 1, male). The complete (not collapsed) list of categories included a *voluntary response*, when an instructor posed a question and an individual raised their hand to answer without conferring with their group; an *individual spontaneous question*, in which a student asked an instructor an unprompted question or was only very generally prompted (e.g., "Does anyone have a question?"); an *individual spontaneous call*, when a student made a comment not prompted by the instructor; a *cold call*, a nonvoluntary response after the instructor called randomly on an individual (in this scenario, the students did not confer with a group); a *spontaneous call post–think–pair–share (TPS)*, a nonvoluntary response after the instructor called randomly on a group after that group had discussed a posed question; a *voluntary response post-TPS*, a voluntary response after the instructor posed a question, the students conferred, and a student volunteered to answer the question; a *voluntary response post-TPS and clicker*, a voluntary response after the instructor posed a question, the students conferred, the students answered the question using a personal response system (e.g., iclicker, TopHat, ChimeIn), and

Table 2. Six universities participated in the current study, representing diverse geographic locations across the world.

Institution	Location	Undergraduate enrollment	Institution type	Number of courses sampled
American University in Cairo	Cairo, Egypt	5474	Private	4
Bethel University	St Paul, Minnesota	2800	Faith-based, private	1
Cornell University	Ithaca, New York	14,907	Public and private	2
University of Bergen	Bergen, Norway	17,000	Public	2
University of Minnesota	Minneapolis, Minnesota	30,511	Public	32
University of Puget Sound	Tacoma, Washington	2553	Private	3

then a student volunteered to answer the question (either after or before the instructor showed the answer; this category is different from voluntary response post-TPS in that the students had committed to an answer before responding); and a *circulating instructor question or comment*, when the instructor circulated around the classroom and a student called them over with a question or comment. (We do not distinguish on the basis of the content of the interaction, because it is often difficult to identify what is said from the observer's perspective.)

To increase the power of our analyses, we focused on the most robust categories or combined relevant values to create broader categories. The final values we included in the analyses were voluntary responses, the most common type of interaction in which an instructor posed a question and an individual raised their hand to answer without conferring with their group; group responses, or any interactions that occurred between the student and the instructor after the students had had some opportunity to discuss a topic with their group members; and total responses, or all interactions between the student and the instructor. To clarify, although the total responses category is not exclusive to voluntary or group responses, voluntary and group responses are exclusive to one another. The total responses category is the sum of the voluntary and group responses, in addition to a small number of additional interactions from the original categories described above. Across the 2 years of observations, the interobserver reliability at UNM was consistently well within an acceptable range for the observers' ability to identify voluntary responses and group responses (Cohen's kappa > .90; Hallgren 2012).

Because some interactions in our observations were not strictly content related (e.g., the instructor and a student discussed a current event not related to the class) or were used only a few times across all observations, the following categories were excluded from our analysis (but note they were included in the total responses variable): individual spontaneous questions, individual spontaneous calls, cold calls, and circulating instructor questions or comments. For example, students asked individual spontaneous questions in the beginning of class more often than at any other point during a class session, and these rarely related to the material. Instead, we prioritized the following categories, because they reliably produced content-related interactions between the instructors and the students: voluntary responses, spontaneous calls post-TPS, voluntary responses post-TPS, and voluntary responses post-TPS and clicker. We included courses with at least two full-class observations (with a minimum of 2, a maximum of 20, and an average of 9.6 observations per course). Only categories that had a total of five or more student–instructor interactions across the observed class sessions for a given course were included in the analyses.

Quantifying predictor variables. To measure the abundance of instructor–student interactions in class, we calculated the average number of student–instructor interactions per class period across all of the observed class periods. The class period duration varied, so when appropriate, we scaled the average number of interactions to fit a 50-minute class period. To measure the diversity of these interactions, we applied Simpson's diversity index to calculate the equitability, or evenness, of teaching strategies per class (Simpson 1949).

Classically, Simpson's diversity index is calculated using the number and abundance of biological species observed and is used in ecology to quantify the biodiversity within a habitat. By considering relative abundances, a diversity index depends not only on species richness but on the evenness of individuals distributed among species. In the present article, we used the number of interaction types and how often instructors used each interaction type to quantify Simpson's diversity index of teaching strategies within a classroom (see supplement 1 for details and an equation). The values range from 0 to 1, with 1 being complete evenness of teaching strategies. In an education context, low values reflect classrooms with little variation in instructor–student interaction types; high values reflect classrooms with lots of different types of instructor–student interactions.

We measured the proportion of women in the class using institutional data when possible and using information from survey data obtained at the beginning of the semester that asked, "Which pronoun do you prefer to describe yourself?" The students could choose among *she/her*, *he/him*, *they/them*, or *other*. The instructors' gender was determined at three levels: man (or men), woman (or women), and both (both men and women). This is because some classes were taught by a man or woman or cotaught by men only, women

Table 3. Best fit models for analyses of total responses, voluntary response, and group response across all institutions.

Outcome variable	Best fit model
Total responses	Approximate class size + (1 per university)
Voluntary response	Approximate class size + (1 per university)
Group response	Approximate class size + Simpson's diversity index + (1 per university)

only, or both men and women, for which we obtained measurements for each instructor. We obtained class size information from the institution or directly from the instructor.

We categorized the classes at two levels—those that primarily enrolled first and second year students (lower division) and those that enrolled third and fourth year students (upper division). We acknowledge that the students in upper division courses do not represent a random sample of students from lower division courses; multiple selective forces may have shaped the student samples.

Statistical analyses

We measured outcomes as likelihood ratios, LR_W, or the likelihood that a participant was a woman compared with the likelihood that a participant was a man in a given category of interaction, such that a value of 1 means that the likelihood of a woman participating is the same as that of a man participating. To calculate the likelihood ratios, we divided the proportion of instructor–student interactions with women, I_w, by the proportion of women in the class, C_w. We then took this value and divided it by the proportion of instructor–student interactions with men, I_m, in turn divided by the proportion of men in the class, C_m:

$$\mathrm{LR_w = (I_w/C_w)/(I_m/C_m)}$$

For example, consider a semester over which we observed student participation in one class. We found that, of all of the student–instructor interactions observed, 30% involved female students, and 70% involved male students. In this example, the class composition was 80 women and 120 men (in other words, 40% women and 60% men). With these values, our outcome would be ((.30/.40)/(.70/.60)) = .64 (i.e., in this class, women participated .64 times as much as men participated). Values less than 1 indicate that women were less likely to participate than men, and values above one indicate that women were more likely to participate. We used linear mixed-effects models with the LME4 package in R (Bates et al. 2014, R Core Team 2014) to test the impact of predictors on the following outcome percentage differentials across institutions: voluntary responses, group responses, and total responses. We used the number of classroom observations as a weighted variable, because it encodes how many original observations were conducted in each classroom, and therefore, larger weights were assigned to courses with more reliable estimates. A model that treated all of the classroom data sets equally would give less observed classes more influence and highly observed classes too little influence. Weighting variables gives each data point the appropriate amount of influence over the parameter estimates and is particularly useful in smaller data sets.

For the multiuniversity analyses, we included schools as a random variable in the mixed-effects model. Starting with a null model, we used Akaike's information criterion (AIC) to assess the model's fit (table 3). We chose the most parsimonious model that best fit the data by calculating AIC differences (Δi) and Akaike weights (w_i), which represent different ways to assess the strength of each model as the best model. We included only data that included all predictor variables (supplement 2).

Because the majority of the classes observed were at UMN, we were also interested in whether apparent trends persisted across the non-UMN institutions (n = 12). We ran post hoc analyses on non-UMN institutions to address this question.

Analyses of courses across six universities with mixed-effects models

Overall, across all of the classes, the average likelihood ratio for voluntary, group, and total interactions were 1.03 (standard deviation [SD] = 0.92), 0.86 (SD = 0.81), and 1.2 (SD = 0.91), respectively. To examine factors that explain observed variation in the data, we used linear mixed-effects models across the 44 classes. Our multilevel model accounted for fixed and random effects to explain variation in the data (e.g., instructor gender as a fixed effect, and school as a random effect). This approach controls for the nonindependence in sampling due to the nested nature of our data (Theobald 2018).

We present data to falsify a number of alternative hypotheses: In our sample of observed classes, gender bias in participation was not predicted by the abundance of interactions in the class (supplement 1), the genders of the instructors (figure 1a), the proportion of women sitting in the classroom (i.e., the critical mass effect; figure 1b), or whether the courses were lower (first and second year) or upper division (third or fourth year; figure 2a).

During the model selection process, all of these variables were eliminated, because they did not significantly improve the fit of the model to the data (supplement 2). The classroom trait that had the largest impact on equitable participation was class size, with women demonstrating higher levels of voluntary responses and total responses in smaller classes across six institutions (voluntary responses, B = –.005, t(24.810) = –3.483, p = .002, standard error [SE] = 0.001; total responses, B = –.004, t(25.274) = –2.890 p = .008, SE = 0.001; figure 3). According to these estimates, as class size increased, fewer women were likely to voluntarily respond to questions posed by the instructor. On the basis of the estimated effect size, an increase in class size from 50 to 150 students decreased the likelihood

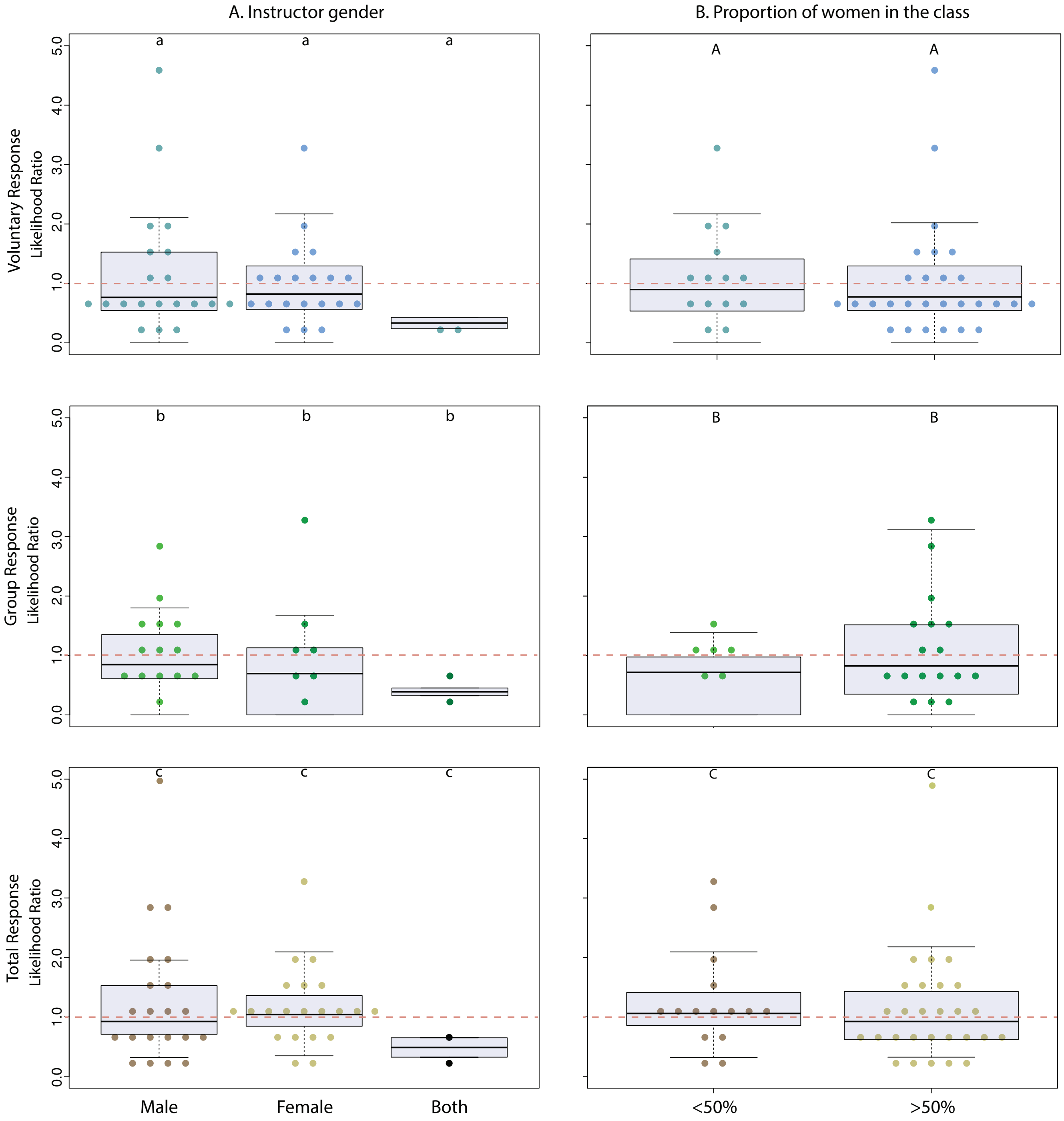

***Figure 1.** (a) Instructor gender: The likelihood of female voluntary responses, group responses, and total responses based on the instructor's gender. (b) The proportion of women in the classroom: The likelihood of female voluntary responses, group responses, and total responses based on the proportion of women in the classroom (either under 50% or over 50%). The letters at the top of each panel indicate insignificant differences (p > .05). Values less than 1 indicate fewer women participated relative to men, and values above 1 indicate that more women participated. The dashed line indicates parity in participation.*

of a woman participating relative to a man by 50%. Class size did not have a significant impact on gender-specific group responses across the six institutions (*B* = –.004, *t*(17.805) = –1.643, *p* = .118, SE = 0.002). The Simpson's diversity index, which considers the variety of interactions and how often instructors used each type of interaction, significantly predicted the group response likelihood ratios (*B* = 2.114, *t*(26.897) = 2.473, *p* = .020, SE = 0.855; figure 4a), with increasing likelihood of female participation as the teaching methods varied. Future research will profit from an

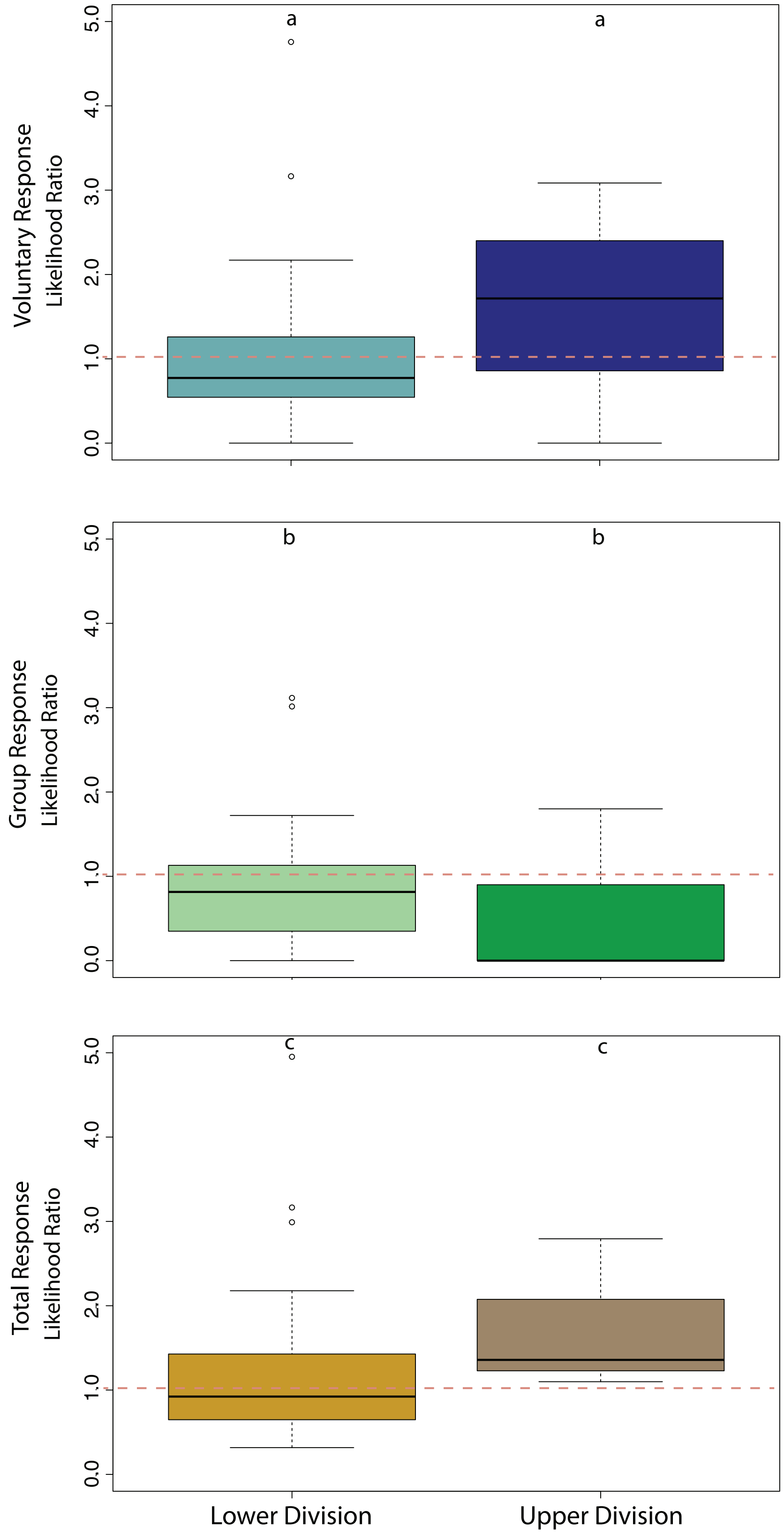

*Figure 2. The likelihood of female voluntary responses, group responses, and total responses in lower division versus upper division courses across all institutions. The letters above the box plots show a lack of statistical significance across categories (*p *> .05).*

explicit focus on this course component to clarify the full impact of group discussions on equitable participation.

In order to test whether the relationship between class size and the likelihood of woman participation was driven by the data obtained from UMN, we combined and analyzed all of the institutions other than UMN. Because of the low sample size (n = 12), we caution readers as they interpret our results. Using Spearman's rank-order correlations, we found significant negative correlations between class size and the likelihood of female participation with voluntary responses (r_S = –.774, p = .003) and total responses (r_S = –.770, p = .003) but not with group responses (n = 9, r_S = –.200, p = .606) across the 12 non-UMN classes (supplement 3). For the Simpson's diversity index, we did not observe the same results when we removed UMN. We found significant negative correlations between Simpson's diversity index and the likelihood of female participation across voluntary responses (r_S = –.755, p = .005) and total responses (r_S = –.664, p = .018) but not across group responses (n = 9, r_S = –.050, p = .898; supplement 3).

Predictors of equitable participation

We analyzed predictors of female participation as voluntary responses, group responses, and total responses in a class session across 44 unique STEM courses (table 4). We falsified several alternative hypotheses and demonstrated that gender-biased participation sharply increases in large classes. These results suggest that the reluctance of women to participate in class is related to traits inherent to large classrooms. We also used a modified form of Simpson's diversity index and equitability as a proxy for diverse teaching strategies in student–instructor interactions (described in supplement 1). The Simpson's diversity index measure showed that women were more likely to participate after group work when the instructor employed diverse teaching strategies in the course.

The impacts of class size

Research on the reduction of class size has produced mixed results and has

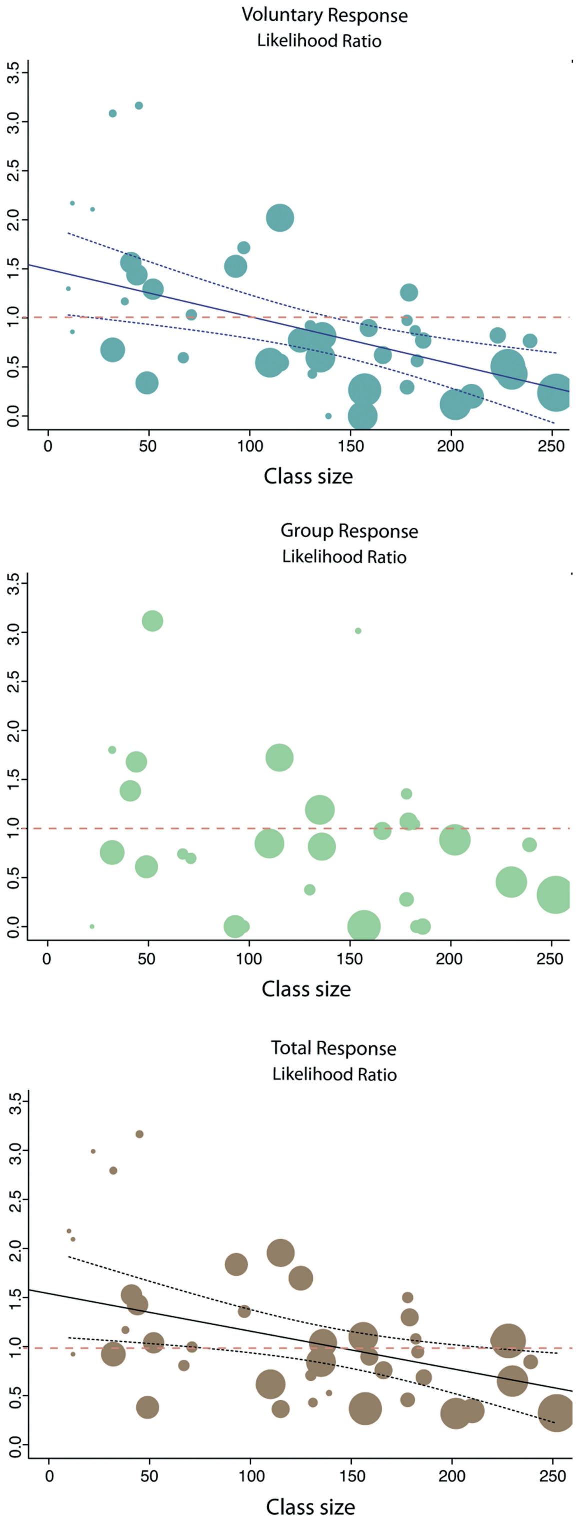

Figure 3. The impact of class size on the likelihood of female voluntary responses, group responses, and total responses across all institutions sampled. The regression lines with confidence intervals denote significant relationships between the likelihood ratio and class size (p < .05), with values below 1 indicating that women were less likely to participate than men. The size of the symbol is proportional to the number of classes observed.

been largely focused on K–12 student populations and used much smaller scales than the data presented in the present article. Despite ongoing debates on the effectiveness of reducing class size in K–12 learning spaces, several state legislatures have appropriated significant amounts of money to reduce classes to between 15 and 20 students (summarized in Zinth 2005). For example, in 1990, the Tennessee legislature funded a longitudinal study on the impact of reducing the size of K–3 classes on student achievement. By following 7000 students across 79 elementary schools, the researchers concluded that small class sizes (13–17 students) increased student achievement scores relative to students in regular class sizes (22–25 students). Furthermore, those students who were exposed to small classes early in their education excelled later, after they were reintroduced into regular-size classes.

Inspired by the results observed in Tennessee, California passed an ambitious education reform initiative in 1996, committing more than $1 billion a year to a class-size reduction program that provided irresistible financial incentives to school districts that reduced the number of students in K–3 classes. However, California schools confronted unique problems that did not apply in the Tennessee case study, including a shortage of qualified teachers and adequate teaching facilities to reduce class size. In addition, California was more culturally diverse, with one-third of California's students living in households in which languages other than English were primarily spoken. Research into California's efforts found that class-size reduction did not benefit school districts serving the state's most historically underserved students. This was partly because the effort was more expensive to implement than expected; in efforts to recruit new staff, they observed a decline in average teacher qualifications, and in order to create additional classroom spaces, lower-income schools used facilities and resources at the expense of other programs (Jepsen and Rivkin 2009). Therefore, the impacts of class size reduction efforts can be context dependent, and care must be taken in assessing them.

The results from studies that were focused on the effects of class size in higher education approach the research on a different scale and generally with more diverse student populations. Cuseo (2007) reviewed studies in which the effects of class size on teaching, learning, and retention were examined. His findings indicate that increasing class size had deleterious impacts on the educational outcomes of students overall and of students enrolled in first-year courses in particular. Studies using big data have echoed these findings—that student achievement declines as class size increases (Dillon et al. 2002, Kokkelenberg et al. 2008). Maringe and Sing (2014) warned that increasing class sizes are particularly dangerous when coupled with current national trends toward increased student mobility, access to higher education, and internationalization of student composition. They point to the impact of the

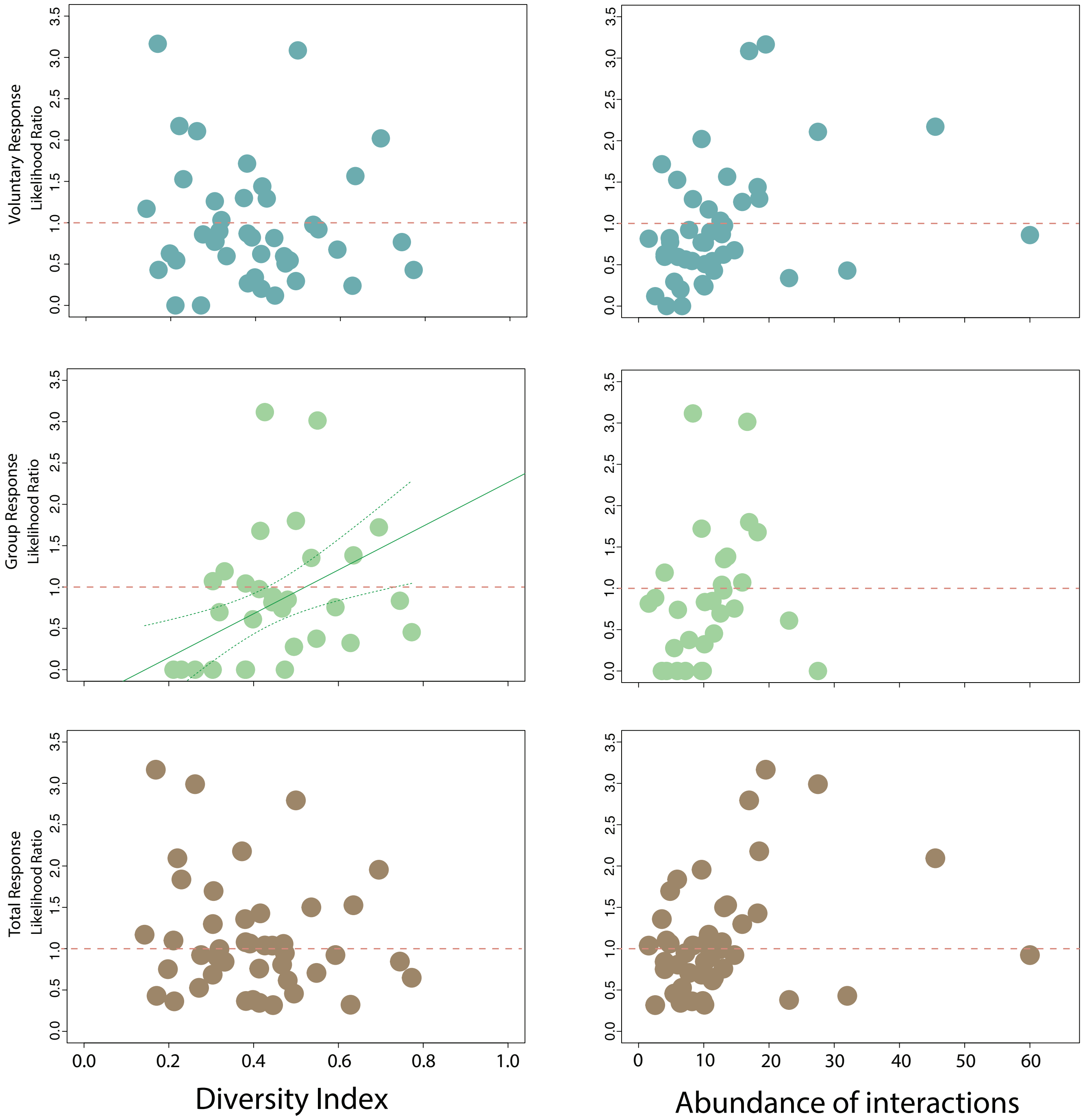

***Figure 4.** The likelihood of female voluntary responses, group responses, and total responses across all institutions as a function of a calculated in-class Simpson's diversity index that measured the amount of varied teaching strategies an instructor used and the abundance of interactions per 50-minute class period. Regardless of class size, more women participated after group discussions when the instructor used more diverse types of interactions during the class period. The regression lines with confidence intervals denote significant relationships between variables (p < .05), with values below 1 indicating women were less likely to participate than men.*

trade-off between individualized instruction and class size on student participation and engagement, curricular access and interpretation, opportunities for deep learning for all, and evaluation of student learning and satisfaction.

Renewed focus on this topic is warranted after the recent development of online or hybrid classes and very large enrollments. For example, students in the University of Central Florida's College of Business obtained more than 1800 signatures on a petition criticizing the college's

Table 4. Summary of results found in the observational study of student participation across six institutions.

Course element tested	Difference?	Notes
Abundance of student–instructor interactions	No	No effect
Diversity of student–instructor interactions	Yes	More diverse interactions = more female participation after group work
Proportion of women in the classroom	No	No effect
Instructor gender	No	No effect
Class size	Yes	Smaller class size = more female participation in voluntary responses and across all observations
Lower or upper division course	No	No effect

recent shift to a blended classroom model. Classes that tend to have between 800 and 2000 students learn through a reduced class time format, which eliminates lectures with the expectation that students spend more time learning with their peers outside of class to gain more thorough knowledge of the material (*www.insidehighered.com/digital-learning/article/2018/09/21/blended-learning-model-university-central-florida-draws-business*). From an institutional perspective, although the additional costs of smaller classes are viewed as prohibitively expensive as enrollment rises, results such as those of the present study should not be ignored. Increased understanding of the qualities that support learning and participation of students in small, medium, and large classes will improve those courses' effectiveness.

Why do we observe gender differences in participation?

Our data show that the largest gender disparities in participation occurred when the instructors elicited voluntary responses from students immediately after asking a question in a large lecture hall. Previous work suggests that instructors may not provide enough time for most students to think through a response. Rowe (1974a) reported that, when precollege instructors asked voluntary response questions, the wait time before the instructor rephrased or called on a student was approximately 1 second. With approximately 1 second, students must formulate a response and decide whether to participate, and many factors unrelated to content knowledge affect the decision to do so. Some of these factors may differentially affect men and women. For example, Cooper and colleagues (2018) showed that men generally have a higher perception of their own ability in a disciplinary domain. In the context of an interactive introductory STEM course, this may lead to increased comfort among men in readily participating in front of a large lecture.

Other work shows that different factors prevent men and women from participating, with women citing a central reason as "not working up the nerve" to ask a question or respond to an answer (Ballen et al. 2018b, Carter et al. 2017). Elements of social identity threat may also be at work, in which a person's social identity (in this case, gender), can be—or are perceived to be—negatively stereotyped (Steele et al. 2002). Evidence from the precollege literature suggests that, regardless of how girls perform in a subject, they are more concerned about how the instructor will evaluate them (Pomerantz et al. 2002) and are less confident than boys in their science content knowledge, even after controlling for variation in their performance (Micari et al. 2007). According to Micari and colleagues, this difference is apparent in several STEM disciplines at the college level and likely plays a role in the observed skewed in-class participation toward males.

Limitations

The methods of this study have a number of limitations. We decided to quantify real-time interactions in classrooms to expand our opportunities to collaborate across universities. However, this meant that, in some classes, the observers could not double check whether they had categorized interactions correctly if they were unsure. An advantage of having real-time observers in the classroom is a reduced uncertainty about student gender, and the observers could move if necessary to better identify students (which is not possible with a camera). Although the person who trained all of the observers was the same (CJB), we were only able to obtain reliability scores across observers at UMN. Within the categories we used (voluntary response or response after group work) we consistently had very high interobserver reliability at UNM (above .90), but this was not measured across all of the observers. Therefore, we cannot rule out the possibility that the reliability across other institutions was lower than that at UMN. However, for this reason, we urge readers to find the analyses of total responses the most reliable, because they encompass all types of interactions. In addition, for responses for which the instructor posed a question and selected a person to answer, there is the possibility that the instructor, being aware of the ongoing study, would preferentially select women more often. The instructors reported that they did not knowingly do this, and the results were similar between individual spontaneous questions (i.e., those in which a student asked an instructor an unprompted question or was only very generally prompted), where this was not an issue, and the other categories.

Another limitation is the binary assignment of gender. Such assignment may not align with self-identified gender. Gender does not exist as a binary variable but rather along a continuum (Ainsworth 2015). In the present study, we only reported male and female genders because of the limitations of our noninvasive observation methods, and we recognize that we are unable to report more accurate gender identities.

Although we focused on either lower division (first and second year) or upper division (third or fourth year) classes, this does not rule out the possibility that the course level precisely reflects the composition of student experience in those courses. Specifically, some introductory classes that are required for certain majors can be taken at any time before graduation and might include larger proportions of older students than other introductory classes. We did not examine the composition of students in those classes in this context specifically.

Finally, we removed one class from the analysis, because it yielded an unusually high likelihood ratio. Whereas all other values ranged from 0 to 4 (i.e., the likelihood of female participation was up to four times that of male students), in this class, the likelihood of women participating was 18 times higher in two types of participation. We believe this may have been the impact of one or two very vocal students. Although the outlier did not affect the overall results, it created significant relationships between participation and class division. Because we cannot completely rule out the possibility that the results that include this data point are a better explanation of student participation in science, we also provide the model selection and results as they appear with the inclusion of this outlier (supplement 2). Although the current data set has limitations, this kind of collaborative effort among universities still allows us to amass enough data to assess predictors of behavior and answer larger questions across a broad sample of university types.

What can instructors do to broaden participation?

Instructors who teach large lectures can use many simple, evidence-based strategies to increase participation. For instance, by simply lengthening wait time after asking a question from 1 second to between 3 and 5 seconds, Rowe (1974b) found that more students volunteered answers and that the students' answers were longer and more complex. In addition, asking students to discuss questions in pairs or in groups lets them work through problems in a nonthreatening environment and practice expressing their opinions prior to being called on (Smith et al. 2009). Our results show that group work mitigated the negative impact of large class size on female participation. Interdependency theory (Rusbult and Van Lange 2008) predicts that individuals who are put in positions to invest in and rely on peers for their success will also help themselves. Previous work demonstrates how increasing interdependency among classroom peers promotes participation, discussion, and ideas (Brewer and Klein 2006). In large classrooms, structured ways to promote interdependency among students is one pathway to improve equitable participation.

Another simple option is to have students respond in writing first rather than out loud, using a student response system that has space for open responses to questions. After the instructor reports a few anonymous notable answers, they can ask the students to follow up out loud. To increase the breadth of responses in class, instructors can ask for multiple volunteers and only call on one or more individuals after a certain number of students have raised their hands (Tanner 2013). Instructors can assign student groups a number and can use a random number generator to spontaneously call on the groups. Within student groups, randomly appointed reporters can be responsible for voicing an answer on behalf of their group, which also takes the responsibility off of the individual if the answer is incorrect (Cohen and Lotan 2014). Instructors can assign reporters on the basis of arbitrary qualities, such as the person who woke up earliest that morning or the person sitting closest to the classroom entryway (Tanner et al. 2013). Critically, our findings suggest that employing a diversity of strategies to promote engagement, rather than simply settling on one or two, is likely to lead to more equitable participation.

We did not explicitly address engagement in this research, but future research will profit from the study of engagement equity as a function of class size. If women are experiencing large classes differently from men, which contributes to gender gaps in participation, we may also expect differences in engagement as well.

For students, the opportunity to reflect on, interact with, and come to a deep understanding of scientific ideas is central to learning. Providing explicit guidance for instructors requires a careful investigation of the underlying factors that contribute to observed classroom disparities.

Conclusions

Our results align with previous work that has called for a halt to the continued expansion of large, introductory gateway courses in science (Cuseo 2007, Achilles 2012, Baker et al. 2016) and underscores the importance of continued empirical measurement of factors that either promote or counter equity in undergraduate STEM (Brewer and Smith 2011, National Academies of Sciences and Medicine 2016). In practice, the gender gap in participation means that women in large STEM courses systematically miss out on opportunities to rehearse articulating their answers aloud to a science community in an environment where wrong answers rarely have negative impacts on consequential outcomes, such as grades. These formative experiences are bound to influence future interactions (e.g., in seminars and conferences; Pritchard et al. 2014, Carter et al. 2017, Hinsley et al. 2017, Schmidt and Davenport 2017, Schmidt et al. 2017), possibly contributing to a general tendency to undervalue the input of women in STEM (e.g., as grant recipients or speakers; Isbell et al. 2012, Grunspan et al. 2016).

Fortunately, although large lectures do pose a clear challenge to student success overall, and to equitable performance (Ballen et al. 2018) and participation specifically, instructors can employ simple strategies to minimize some of these challenges. In fact, many evidence-based active-learning techniques appear to work by making large classes function like smaller classes. Our results show that women were more likely to participate after small group discussions, and this effect was more pronounced when diverse teaching approaches were employed. Furthermore, these findings support the course deficit model, whereby overt instructional choices can minimize gaps—in this case, in participation—that may contribute to inequalities in STEM (Cotner and Ballen 2017). By placing some of the burden of responsibility on instructors, we are in a better position to be proactive in our classrooms with respect to these inequities.

We realize that, ultimately, administrators and legislators must grapple with the problems associated with large classes, and we hope this work can be part of that conversation. On the basis of our results, large classes begin to negatively affect students when they include more than approximately 120 students. This may be because class size is strongly associated with the kinds of assignments given and the level of student involvement in class. Instructors can play an active role in minimizing the problems associated with large classes by drawing on the active-learning literature and exploring which strategies, from an array of possibilities, are most effective in their own courses. Our results suggest that the best way to ameliorate the negative impact of large class sizes on female participation is to use diverse teaching strategies and small group interactions.

Acknowledgments

This work would not have been possible without undergraduate and graduate research assistants, including Neelam Chandiramani, Melissa Khaw, Shivonne McCarthy, Sergio Molina, Jake Peterson, Meredith Song, Morgan Burkhart, Grace Leland, Brandon Vanderbush, Mai Vang, Christine Lian, Connor Neil, Tiare Gill, Kylie Young, Irene Deng, Ellica Spiut, and Gregor-Fausto Siegmund, for collecting and entering participation data; Lucy Fortson and Mos Kaveh, for connecting us with instructors; Jennifer Kroschel, Geri Grosinger, Oddfrid T. Kårstad Førland, and Rachael Shelden, for invaluable logistical support; Aaron Wynveen, Annika Moe, Mark Decker, and Cheryl Scott, for allowing us to observe their classes; and J.D. Walker, for statistical advice and valuable insights. This work was funded in part by a Research Coordination Network grant no.1729935, awarded to SC and CJB, from the National Science Foundation, RCN–UBE Incubator: Equity and Diversity in Undergraduate STEM.

Supplemental material

Supplemental data are available at *BIOSCI* online.

References cited

Achilles CM. 2012. Class-Size Policy: The STAR Experiment and Related Class-Size Studies. National Council of Professors of Educational Administration. Policy brief vol. 1, no. 2.

Ainsworth C. 2015. Sex redefined. Nature 518: 288.

Baker BD, Farrie D, Sciarra DG. 2016. Mind the gap: 20 years of progress and retrenchment in school funding and achievement gaps. ETS Research Report Series 2016: 1–37.

Ballen CJ, Aguillon SM, Brunelli R, Drake AG, Wassenberg D, Weiss SL, Zamudio KR, Cotner S. 2018a. Do small classes in higher education reduce performance gaps in STEM? BioScience 68: 593–600.

Ballen CJ, Lee D, Rakner L, Cotner S. 2018b. Politics a “chilly” environment for undergraduate women in Norway. PS: Political Science and Politics 51: 653–658.

Ballen CJ, Wieman C, Salehi S, Searle JB, Zamudio KR. 2017. Enhancing diversity in undergraduate science: Self-efficacy drives performance gains with active learning. CBE—Life Sciences Education 16 (art. 56).

Bates D, Mächler M, Bolker B, Walker S. 2014. Fitting linear mixed-effects models using lme4. ArXiv. Cornel University. https://arxiv.org/abs/1406.5823.

Beede DN, Julian TA, Langdon D, McKittrick G, Khan B, Doms ME. 2011. Women in STEM: A gender gap to innovation. US Department of Commerce. Economics and Statistics Administration issue brief no. 04-11.

Beichner RJ, Saul JM, Abbott DS, Morse JJ, Deardorff D, Allain RJ, Bonham SW, Dancy MH, Risley JS. 2007. The student-centered activities for large enrollment undergraduate programs (SCALE-UP) project. Research-Based Reform of University Physics 1: 2–39.

Bornmann L, Mutz R, Daniel H-D. 2007. Gender differences in grant peer review: A meta-analysis. Journal of Informetrics 1: 226–238.

Brewer CA, Smith D. 2011. Vision and Change in Undergraduate Biology Education: A Call to Action. American Association for the Advancement of Science.

Brewer S, Klein JD. 2006. Type of positive interdependence and affiliation motive in an asynchronous, collaborative learning environment. Educational Technology Research and Development 54: 331–354.

Carter A, Croft A, Lukas D, Sandstrom G. 2017. Women’s visibility in academic seminars: Women ask fewer questions than men. ArXiv. Cornell University. https://arxiv.org/abs/1711.10985.

Cohen EG, Lotan RA. 2014. Designing Groupwork: Strategies for the Heterogeneous Classroom, 3rd ed. Teachers College Press.

Cooper KM, Krieg A, Brownell SE. 2018. Who perceives they are smarter? Exploring the influence of student characteristics on student academic self-concept in physiology. Advances in Physiology Education 42: 200–208.

Cotner S, Ballen CJ. 2017. Can mixed assessment methods make biology classes more equitable? PLOS ONE 12 (art. e0189610).

Cotner S, Ballen C, Brooks DC, Moore R. 2011. Instructor gender and student confidence in the sciences: A need for more role models. Journal of College Science Teaching 40: 96–101.

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. Journal of Higher Education 74: 51–76.

Cuseo J. 2007. The empirical case against large class size: Adverse effects on the teaching, learning, and retention of first-year students. Journal of Faculty Development 21: 5–21.

Dahlerup D. 1988. From a small to a large minority: Women in Scandinavian politics. Scandinavian Political Studies 11: 275–298.

Dillon M, Kokkelenberg EC, Christy SM. 2002. The effects of class size on student achievement in higher education: Applying an earnings function. DigitalCommons@IRL, Cornell University. http://digitalcommons.ilr.cornell.edu/cheri/15.

Eddy SL, Brownell SE, Wenderoth MP. 2014. Gender gaps in achievement and participation in multiple introductory biology classrooms. CBE—Life Sciences Education 13: 478–492.

Freeman S, Eddy SL, McDonough M, Smith MK, Okoroafor N, Jordt H, Wenderoth MP. 2014. Active learning increases student performance in science, engineering, and mathematics. Proceedings of the National Academy of Sciences 111: 8410–8415.

Freeman S, Haak D, Wenderoth MP. 2011. Increased course structure improves performance in introductory biology. CBE Life Sciences Education 10: 175–186.

Grunspan DZ, Eddy SL, Brownell SE, Wiggins BL, Crowe AJ, Goodreau SM. 2016. Males under-estimate academic performance of their female peers in undergraduate biology classrooms. PLOS ONE 11 (art. e0148405).

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.

Hallgren KA. 2012. Computing inter-rater reliability for observational data: An overview and tutorial. Tutorials in Quantitative Methods for Psychology 8: 23.

Hinsley A, Sutherland WJ, Johnston A. 2017. Men ask more questions than women at a scientific conference. PLOS ONE 12 (art. e0185534).

Ho DE, Kelman MG. 2014. Does class size affect the gender gap? A natural experiment in law. Journal of Legal Studies 43: 291–321.

Holman L, Stuart-Fox D, Hauser CE. 2018. The gender gap in science: How long until women are equally represented? PLoS Biology 16: e2004956.

Hoffmann F, Oreopoulos P. 2009. A professor like me the influence of instructor gender on college achievement. Journal of Human Resources 44: 479–494.

Isbell LA, Young TP, Harcourt AH. 2012. Stag parties linger: Continued gender bias in a female-rich scientific discipline. PLOS ONE 7 (art. e49682).

Jepsen C, Rivkin S. 2009. Class size reduction and student achievement the potential tradeoff between teacher quality and class size. Journal of Human Resources 44: 223–250.

Kokkelenberg EC, Dillon M, Christy SM. 2008. The effects of class size on student grades at a public university. Economics of Education Review 27: 221–233.

Komarraju M, Musulkin S, Bhattacharya G. 2010. Role of student–faculty interactions in developing college students' academic self-concept, motivation, and achievement. Journal of College Student Development 51: 332–342.

Kuh GD, Hu S. 2001. The effects of student–faculty interaction in the 1990s. Review of Higher Education 24: 309–332.

Ledin A, Bornmann L, Gannon F, Wallon G. 2007. A persistent problem: Traditional gender roles hold back female scientists. EMBO Reports 8: 982–987.

Lorenzo M, Crouch CH, Mazur E. 2006. Reducing the gender gap in the physics classroom. American Journal of Physics 74: 118–122.

Maringe F, Sing N. 2014. Teaching large classes in an increasingly internationalising higher education environment: Pedagogical, quality and equity issues. Higher Education 67: 761–782.

Micari M, Pazos P, Hartmann MJZ. 2007. A matter of confidence: Gender differences in attitudes toward engaging in lab and course work in undergraduate engineering. Journal of Women and Minorities in Science and Engineering 13: 279–293.

National Academies of Sciences and Medicine E. 2016. Barriers and Opportunities for 2-Year and 4-Year STEM Degrees: Systemic Change to Support Students' Diverse Pathways. National Academies Press.

O'Dorchai S, Meulders D, Crippa F, Margherita A. 2009. She Figures 2009: Statistics and Indicators on Gender Equality in Science. Publications Office of the European Union.

Pomerantz EM, Altermatt ER, Saxon JL. 2002. Making the grade but feeling distressed: Gender differences in academic performance and internal distress. Journal of Educational Psychology 94: 396.

Premo J, Cavagnetto A. 2018. Priming students for whole-class interaction: Using interdependence to support behavioral engagement. Social Psychology of Education 21: 915–935.

Pritchard J, Masters K, Allen J, Contenta F, Huckvale L, Wilkins S, Zocchi A. 2014. Asking gender questions. Astronomy and Geophysics 55: 6–8.

R Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (http://www.R-project.org)

Rowe MB. 1974a. Wait-time and rewards as instructional variables, their influence on language, logic, and fate control. part 1: Wait time. Journal of Research in Science Teaching 11: 81–94.

Rowe MB. 1974b. Relation of wait-time and rewards to the development of language, logic, and fate control, part 2: Rewards. Journal of Research in Science Teaching 11: 291–308.

Rusbult CE, Van Lange PAM. 2008. Why we need interdependence theory. Social and Personality Psychology Compass 2: 2049–2070.

Schanzenbach DW. 2014. Does Class Size Matter? National Education Policy Center.

Schmidt SJ, Davenport JRA. 2017. Who asks questions at astronomy meetings? Nature 1: 1.

Schmidt SJ, Douglas S, Gosnell NM, Muirhead PS, Booth RS, Davenport JRA, Mace GN. 2017. The Role of Gender in Asking Questions at Cool Stars 18 and 19. ArXiv. Cornel University. https://arxiv.org/abs/1704.05260.

Simpson EH. 1949. Measurement of diversity. Nature 163: 688.

Smith MK, Wood WB, Adams WK, Wieman C, Knight JK, Guild N, Su TT. 2009. Why peer discussion improves student performance on in-class concept questions. Science 323: 122 LP-124. http://science.sciencemag.org/content/323/5910/122.abstract

Steele CM, Spencer SJ, Aronson J. 2002. Contending with group image: The psychology of stereotype and social identity threat. Advances in Experimental Social Psychology 34: 379–440.

Tanner KD. 2013. Structure matters: Twenty-one teaching strategies to promote student engagement and cultivate classroom equity. CBE—Life Sciences Education 12: 322–331.

Theobald E. 2018. Students are rarely independent: When, why, and how to use random effects in discipline-based education research. CBE—Life Sciences Education 17: rm2.

Wold A, Wennerás C. 2010. Nepotism and sexism in peer-review. Pages 64–70 in Wyer M, Barbercheck M, Cookmeyer D, Ozturk H, Wayne M, eds. Women, Science, and Technology: A Reader in Feminist Science Studies. Routledge.

Zinth K. 2005. State class-size reduction measures. Education Commission of the States.

Cissy J. Ballen (mjb0100@auburn.edu) is affiliated with the Department of Biological Sciences at Auburn University, in Auburn, Alabama. Cissy J. Ballen, Zoe Koth, Deena Wassenberg, Azariah Yonas, and Sehoya Cotner are affiliated with the Department of Biology Teaching and Learning at the University of Minnesota, in Minneapolis, Minnesota. Stepfanie M. Aguillon, Abby Grace Drake, and Kelly R. Zamudio are affiliated with the Department of Ecology and Evolutionary Biology at Cornell University, in Ithaca, New York. Stepfanie M. Aguillon is affiliated with the Fuller Evolutionary Biology Program, at the Cornell Lab of Ornithology, in Ithaca, New York. Azza Awwad, Aziza Ellozy, Caroline Mitry, and Hoda Mostafa are affiliated with the Center for Learning and Teaching at The American University in Cairo, in Cairo, Egypt. Anne E. Bjune, and Christian Jørgensen are affiliated with the Department of Biological Sciences at the University of Bergen, in Bergen, Norway. Daniel Challou, Steve Jensen, and Shana Watters are affiliated with the Department of Computer Science and Engineering at the University of Minnesota, in Minneapolis, Minnesota. Michelle Driessen is affiliated with the Department of Chemistry at the University of Minnesota, in Minneapolis, Minnesota. Vivian E. Ferry is affiliated with the Department of Chemical Engineering and Materials Science at the University of Minnesota, in Minneapolis, Minnesota. Emma E. Goldberg, Suzanne McGaugh, and William Harcombe are affiliated with the Department of Ecology, Evolution, and Behavior at the University of Minnesota, in Minneapolis, Minnesota. Bryan Mosher is affiliated with the School of Mathematics at the University of Minnesota, in Minneapolis, Minnesota. Renee H. Petipas is affiliated with the Department of Plant Pathology at Washington State University, in Pullman, Washington. Paula A.G. Soneral is affiliated with the Department of Biological Sciences at Bethel University, in Saint Paul, Minnesota. Stacey L. Weiss is affiliated with the Department of Biology at the University of Puget Sound, in Tacoma, Washington.