8-1-2019

Smaller classes promote equitable student participation in STEM

Cissy J. Ballen
Auburn University

Stepfanie M. Aguillon
Cornell University

Azza Awwad
American University in Cairo

Anne E. Bjune
Universitetet i Bergen

Daniel Challou
University of Minnesota Twin Cities

See next page for additional authors

Authors
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

Smaller classes promote equitable student participation in STEM

Cissy J. Ballen[1,2,*], Stepfanie M. Aguillon[3,4], Azza Awwad[5], Anne E. Bjune[6], Daniel Challou[7], Abby Grace Drake[3], Michelle Driessen[8], Aziza Ellozy[5], Vivian E. Ferry[9], Emma E. Goldberg[10], William Harcombe[10], Steve Jensen[7], Christian Jørgensen[6], Zoe Koth[2], Suzanne McGaugh[10], Caroline Mitry[5], Bryan Mosher[11], Hoda Mostafa[5], Renee H. Petipas[12], Paula A.G. Soneral[13], Shana Watters[7], Deena Wassenberg[2], Stacey L. Weiss[14], Azariah Yonas[2], Kelly R. Zamudio[3], Sehoya Cotner[2]

[1]Department of Biological Sciences, Auburn University, Auburn, AL, USA
[2]Department of Biology Teaching and Learning, University of Minnesota, Minneapolis, MN, USA
[3]Department of Ecology & Evolutionary Biology, Cornell University, Ithaca, NY, USA
[4]Fuller Evolutionary Biology Program, Cornell Lab of Ornithology, Ithaca, NY, USA
[5]Center for Learning and Teaching, The American University in Cairo, Cairo, Egypt
[6]Department of Biological Sciences, University of Bergen, Bergen, Norway
[7]Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, USA
[8]Department of Chemistry, University of Minnesota, Minneapolis, MN, USA
[9]Department of Chemical Engineering and Materials Science, University of Minnesota, Minneapolis, MN, USA
[10]Department of Ecology, Evolution and Behavior, University of Minnesota, Minneapolis, MN, USA
[11]School of Mathematics, University of Minnesota, Minneapolis, MN, USA
[12]Department of Plant Pathology, Washington State University, Pullman, WA, USA
[13]Department of Biological Sciences, Bethel University, St. Paul, MN, USA
[14]Department of Biology, University of Puget Sound, Tacoma, WA

This article has been accepted for publication in Bioscience published by Oxford University Press.

Word count: 8,102

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

Abstract

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 based on their numeric proportions. Here we asked which attributes of the learning environment contribute to decreased female participation: abundance of in-class interactions, diversity of interactions, proportion of women in class, instructor gender, 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 5,300 student-instructor interactions observed across multiple institutions. We falsify several alternative hypotheses and demonstrate that increasing class size has the largest negative association. We also found that when instructors use a diverse range of teaching strategies, women are more likely to participate after small-group discussions.

Introduction

Active learning can be distinguished from traditional lecturing through its emphasis on diverse types of engagement strategies, including structured student-instructor interactions during activities or guided inquiry (Haak et al., 2011; Smith et al., 2009). Substantial evidence supports interactive classes as a more effective form of instruction compared to 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.*, 2017b). 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). Thus, it is critical that gateway courses are systematically assessed to identify which elements within the classrooms leads to gaps in participation, and 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 (Carter et al., 2017; Hinsley et al., 2017; Pritchard et al., 2014). 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 & Wenneras, 2010) and grants (Ledin et al., 2007), fewer female first (O'Dorchai et al., 2009) and last authors (Holman et al., 2018; Murray 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). Thus, 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 impact female participation, and used deductive methods to understand each element's relative impact on equitable participation from our sample of observations (Table 1).

We examined how the abundance of interactions, diversity of interactions, instructor gender, proportion of women in the class, class size, and class division affect three specific types of student participation: (1) **voluntary responses**, when an instructor poses a question and an individual raises their hand to answer without conferring with their peers; (2) **group responses**, when an instructor poses a question and students have the opportunity to talk to their peers before answering; (3) **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 lectures 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).

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 under-represented 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.

Data collection

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 from instructors who teach majors, nonmajors, or both, from a range of institutions. Volunteers represent Bethel University, Cornell University, University of Minnesota, University of Puget Sound, the American University in Cairo, Egypt, and University of Bergen, Norway (Table 2). 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 5,200 students enrolled in the sampled courses, and observers categorized over 5,300 interactions between the instructors and students (Research Coordination Network, National Science Foundation RCN–UBE Incubator: Equity and Diversity in Undergraduate STEM; #1729935 awarded to S Cotner and CJ Ballen). We included courses from across STEM fields, including biology, physics, computer science, and chemistry (details in the raw data file). 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 Boards (Bethel IRB 180518; Cornell IRB 1410005010; University of Minnesota IRB 00000800; University of Puget Sound IRB 1617-006; American University in Cairo 2016-2017-0012; University of Bergen NSD 46727).

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

Institution	Location	Undergraduate enrollment	Institution type	# of courses sampled
American University in Cairo	Cairo, Egypt	5,474	Private	4
Bethel University	St Paul, MN, US	2,800	Faith-based, private	1
Cornell University	Ithaca, NY, US	14,907	Public and private	2
University of Bergen	Bergen, Norway	17,000	Public	2
University of Minnesota	Minneapolis, MN, US	30,511	Public	32
University of Puget Sound	Tacoma, WA, US	2,553	Private	3

Research methods

Measuring In-Class Participation

We conducted ~1 hour training sessions for 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 takes place during a class period, an observer recorded the gender of the student participant (1 = male or 0 = female). The complete (not collapsed) list of categories included: (1) 'voluntary response,' when an instructor poses a question, and an individual raises their hand to answer without conferring with their group; (2) 'individual spontaneous question,' in which a student asks an instructor an unprompted question or is only very generally prompted (e.g. 'does anyone have a question?'); (3) 'individual spontaneous call,' when a student makes a comment not prompted by the instructor; (4) 'cold call,' a non-voluntary response after the instructor calls randomly on an individual (in this scenario, students have not conferred with a group); (5) 'spontaneous call post-Think Pair Share (TPS),' a non-voluntary response after the instructor calls randomly on a group after they discuss a posed question; (6) 'voluntary response post-TPS,' a voluntary response after the instructor poses a question, students confer, and a student volunteers to answer the question; (7) 'voluntary response post-TPS and clicker,' a voluntary response after the instructor poses a question, students confer, students answer the question using a personal response system (e.g., iclicker, TopHat, ChimeIn), and then a student volunteers to answer the question (either after the instructor shows the answer or before; this category is different from voluntary response post-TPS (#6) in that students have committed to an answer before responding); and (8) ‘circulating instructor question or comment,’ when the instructor is circulating around the classroom, and a student calls them over with a question or comment (note: we do not distinguish based on content of the interaction because it is often difficult to identify what is said from the observer’s perspective).

To increase power of analyses, we focus on the most robust categories or combined relevant values to create broader categories. The final values we included in analyses were (A) **voluntary responses**, the most common type of interaction in which an instructor poses a question, and an individual raises their hand to answer without conferring with their group (#1 above), and (B) **group responses**, or any interactions that occur between the student and the instructor after students have some opportunity to discuss a topic with group members (combination of #5-7 described above), and (C) **total responses**, or all interactions between the student and instructor. To clarify, while (C) is not exclusive to (A) and (B), (A) and (B) are exclusive to one another. Category (C) is the sum of (A) and (B), in addition to a small number of additional interactions from the original categories described above. Across the two years of observations, inter-observer reliability at the University of Minnesota was consistently well within acceptable range among observers’ ability to identify voluntary responses and group responses (Cohen's kappa > 0.90; Hallgren, 2012).

Because some interactions in our observations were not strictly content related (e.g. instructor and student discuss current event not related to class) or used only a few times across all observations, categories 2-4 and 8 were excluded from our analysis (but note they are included in the total responses variable). For example, students asked individual spontaneous questions in the beginning of class more often than any other point during lecture, and these rarely related to the material. Instead we prioritized categories 1, 5-7 because these reliably produced content-related interactions between instructor and student. We included courses with at least two full-class observations (minimum 2, maximum 20, average 9.6 observations per course).

Only categories that had a total of five or more student–instructor interactions across 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 observed class periods. 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 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 the evenness of individuals distributed among species. Here, we use the number of interaction types, and how often instructors use each interaction type, to quantify Simpson's diversity index of teaching strategies within a classroom (see Supplementary materials 1 for details and equation). 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 used frequently.

We measured the proportion of women in the class using institutional data when possible, and information from survey data obtained at the beginning of the semester that asked "Which pronoun do you prefer to describe yourself?" Students could choose between she/her, he/him, they/them, or other. Instructor gender was estimated at three levels: man (or men), woman (or women), or both (both men and women). This is because some classes were taught **by a man or woman, or** co-taught by men only, women only, or both men and women, for which we obtained measurements from each instructor. We obtained class size information from the institution or directly from the instructor.

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

Statistical analyses

We measured outcomes as likelihood ratios, LR_W, or the likelihood that a participant is a woman compared to the likelihood that a participant is a man in a given category of interaction, such that a value of one means that the likelihood of a woman participating is the same as that of a man. To calculate 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, over proportion of men in the class, C_m.

$$LR_W = (I_w/C_w)/(I_m/C_m)$$

For example, consider a semester over which we observed student participation in one class. We found that of all 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 ((0.30 /0.40)/(0.70/0.60))=0.64 (i.e., in this class women participate 0.64 times as much as men participate). Values less than 1 indicate that women were less likely to participate relative to 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 are 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 datasets.

For the multi-university analyses, we included schools (‘uni’) as a random variable in the mixed effects model. Starting with a null model, we used Akaike's information criterion to assess model 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 both represent different ways to assess of strength of each model as the best model. We only included data that included all predictor variables (Supplementary material 2: Model selection summary tables).

Because the majority of classes observed were from the University of Minnesota (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.

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	~class size + (1|uni)
Voluntary Response	~class size + (1|uni)
Group Response	~class size + Simpson’s diversity index + (1|uni)

Results

Analyses of courses across six universities with mixed-effects models

Overall, across all classes, the average likelihood ratio for voluntary, group, and total interactions were 1.03 (0.92 SD), 0.86 (0.81 SD), and 1.2 (0.91 SD), 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 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 (Supplementary material 1)
- the gender(s) of the instructors (Figure 1A)

- the proportion of women sitting in the classroom (i.e., ‘critical mass effect’; Figure 1B)
- whether 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 (Supplementary material 2: Model selection summary tables; Results tables). 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 = -0.005, t(24.810) = -3.483, P = 0.002, SE = 0.001; total responses B = -0.004, t(25.274) = -2.890 P = 0.008, SE = 0.001; Figure 3). Based on these estimates, as class size increased, fewer women were likely to voluntarily respond to questions posed by the instructor. Based on the estimated effect size, an increase in class size from 50 to 150 students decreased the likelihood of a woman participating relative to a man by 50%. Class size did not have a significant impact on gender-specific group responses across six institutions (B = -0.004, t(17.805) = -1.643, P = 0.118, 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 group response likelihood ratios (B = 2.114, t(26.897) = 2.473, P = 0.020, SE = 0.855; Figure 4A), with increasing likelihood of female participation as teaching methods varied. Future research will profit from an explicit focus on this course component to clarify the full impact of group discussions on equitable participation.

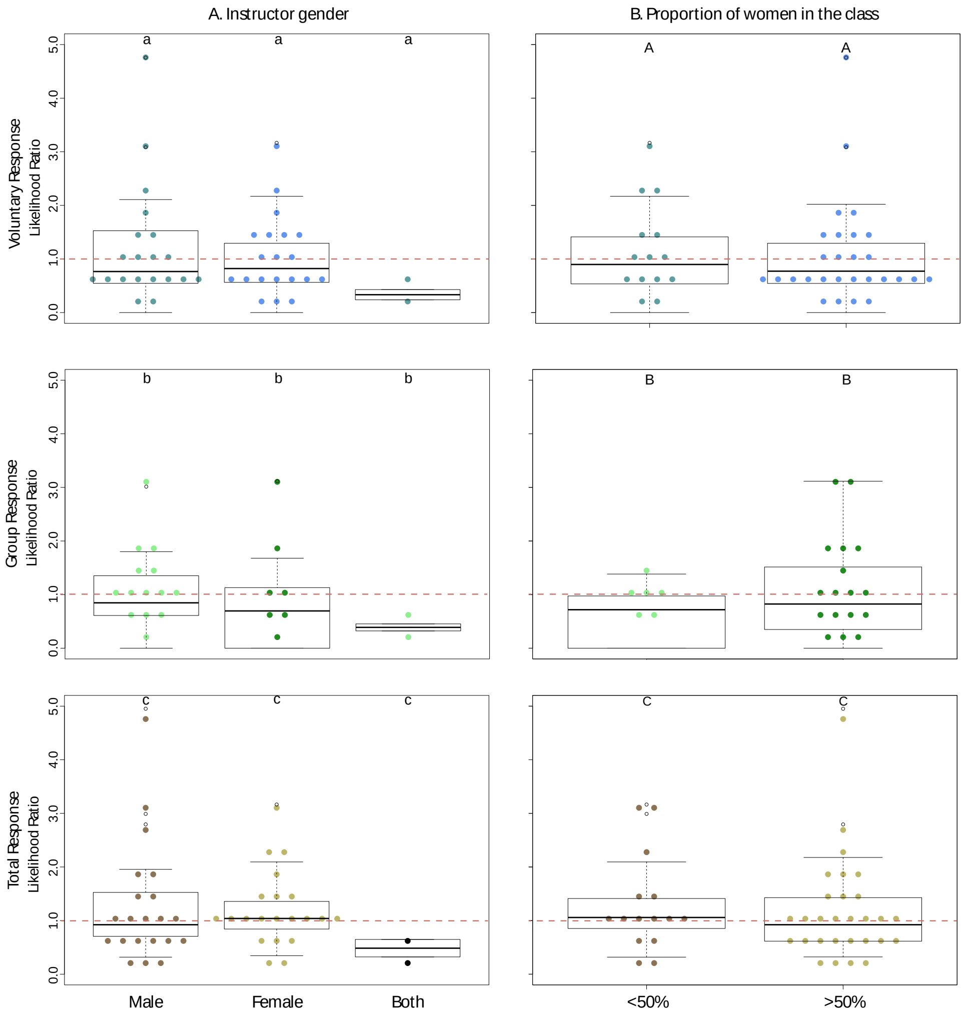

Figure 1. A. Instructor gender: Likelihood of female voluntary responses (blue), group responses (green), and total responses (brown) based on the instructor gender. B.Proportion of women in the classroom: Likelihood of female voluntary responses (blue), group responses (green), and total responses (brown) based on the proportion of women in the classroom (either under 50% or over 50%). Letters at the top of each panel indicate non-significant differences (P>0.05). Values less than 1 indicate fewer women participated relative to men, and values above one indicate more women participated. The dashed line indicates parity in participation.

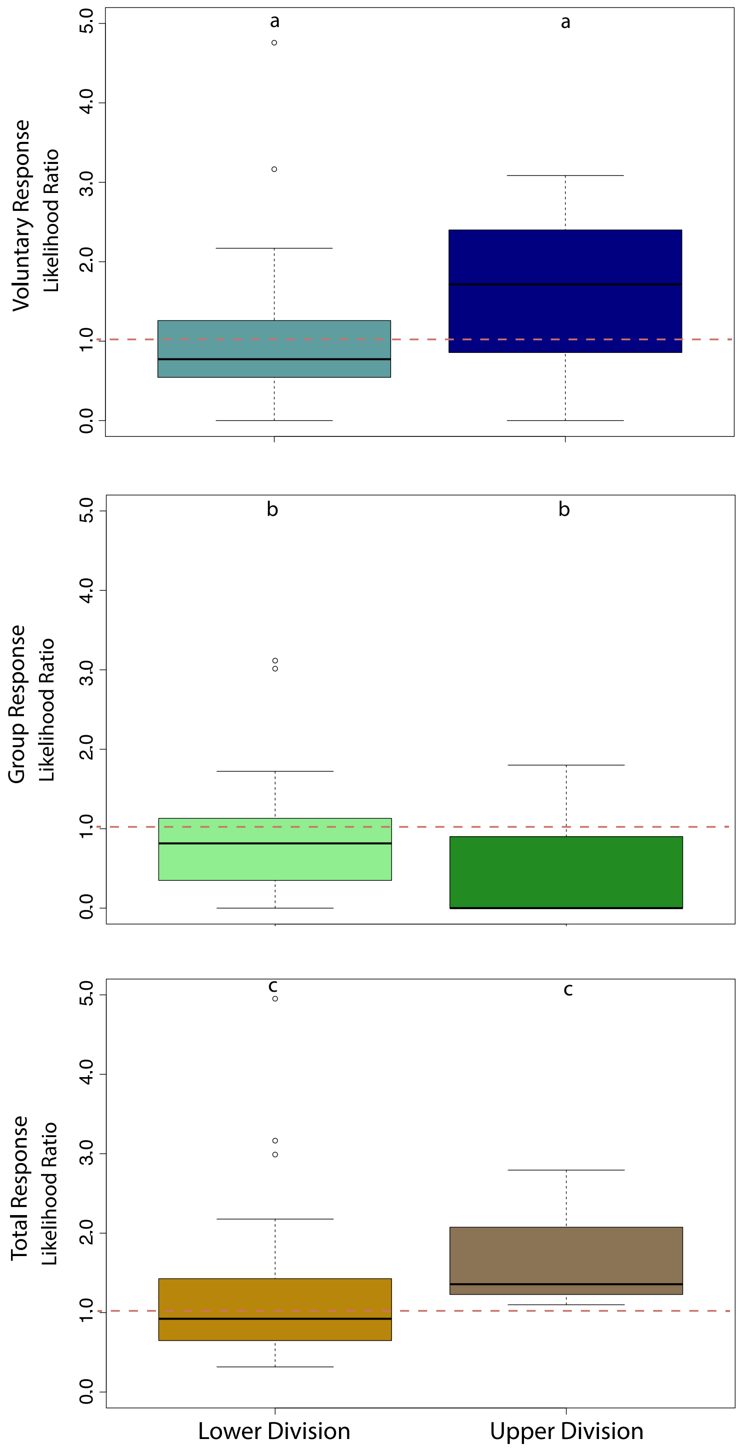

Figure 2. Likelihood of female voluntary responses (blue), group responses (green), and total responses (brown) in lower division versus upper division courses across all institutions. Letters above the box plots show statistical non-significance across categories (P>0.05).

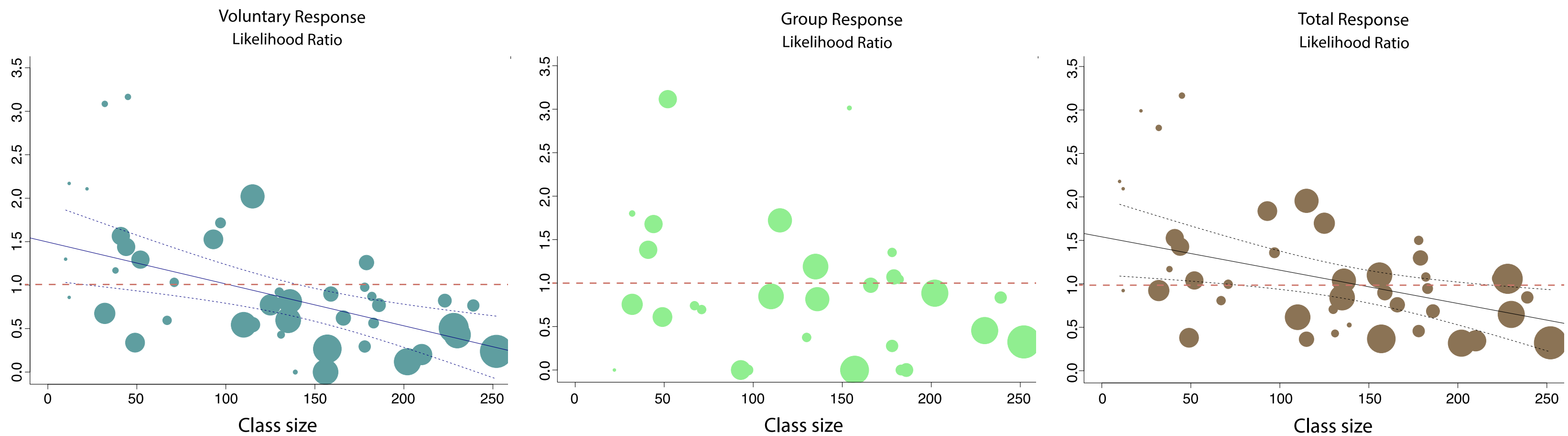

Figure 3. The impact of class size on the likelihood of female voluntary responses (blue), group responses (green), and total responses (brown) across all institutions sampled. Regression lines with confidence intervals denote significant relationships between the likelihood ratio and class size (P<0.05), with values below one indicating women were less likely to participate than men. The size of the symbol is proportional to the number of classes observed.

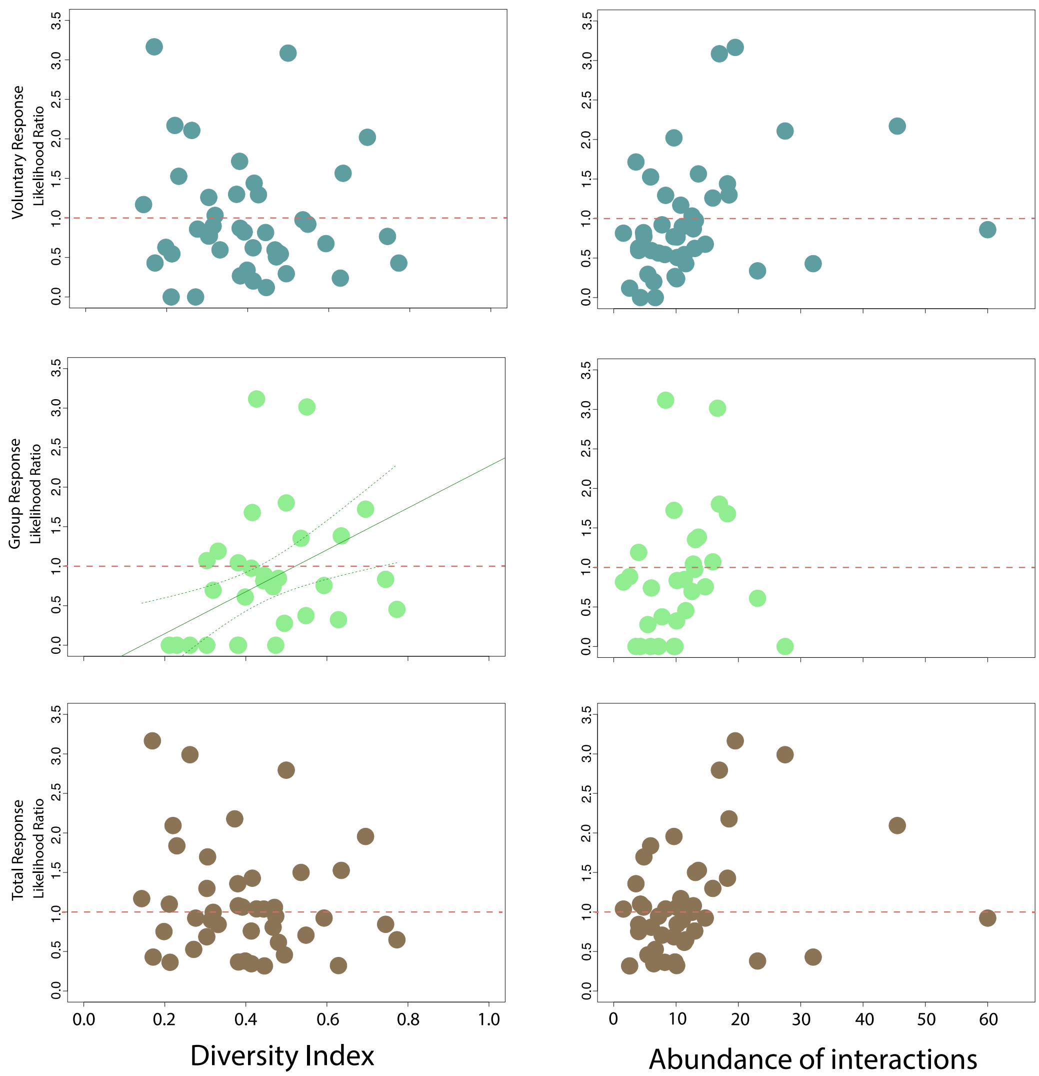

Figure 4. The likelihood of female voluntary responses (blue), group responses (green), and total responses (brown) across all institutions as a function of a calculated in-class ‘Simpson's diversity index’ that measures the amount of varied teaching strategies an instructor uses and the abundance of 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. Regression lines with confidence intervals denote significant relationships between variables ($P<0.05$), with values below one indicating women were less likely to participate than men.

In order to test whether the relationship between class size and likelihood that women participate was driven by the data obtained from the University of Minnesota (UMN), we combined and analyzed all institutions *other than* UMN. Due to the low sample size (N = 12), we caution readers as they interpret our results. Using Spearman's correlations, we found significant negative relationships between class size and the likelihood of female participation with voluntary responses (r_s = -0.774 , P = 0.003) and total responses (r_s = -0.770 , P = 0.003), but not group responses (N = 9; r_s = -0.200, P = 0.606) across the 12 non-UMN classes (Supplementary material 3). For the Simpson's diversity index, we did not observe the same results when we removed University of Minnesota. We found significant negative relationships between Simpson's diversity index and the likelihood of female participation across voluntary responses (r_s = -0.755, P = 0.005) and total responses (r_s = -0.664, P = 0.018), but not group responses (N = 9; r_s = -0.050, P = 0.898; Supplementary material 3).

Discussion

We analyzed predictors of female participation as voluntary responses, group responses, and total responses in lecture, across 44 unique STEM courses (Summary of results, 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 lecture courses. 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 Supplementary materials 1). The Simpson's diversity index measure showed women were more likely to participate after group work when the instructor employed diverse teaching strategies in the course.

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 division or upper division course	No	No effect.

The impacts of class size

Research on the reduction of class size has produced mixed results, largely focused on K-12 student populations, and on much smaller scales than the data presented here. 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 to 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 7,000 students across 79 elementary schools, researchers concluded that small class sizes (13-17 students) increased student achievement scores as compared to students in regular class sizes (22-25 students). Further, those students who were exposed to small classes early in their education excelled later, after they were re-introduced into regular-sized 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. Additionally, 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 (1) the effort was more expensive to implement than expected; (2) in efforts to recruit new staff, they observed a decline in average teacher qualifications; and (3) in order to create additional classroom spaces, lower-income schools used facilities and resources at the expense of other programs (Jepsen and Rivkin, 2009). Thus, impacts of class size reduction efforts can be context-dependent, and care must be taken in assessing their impacts.

Results from studies that focus 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 that examined the effects of class size on teaching, learning and retention. His findings indicate that increasing class size had deleterious impacts on educational outcomes for students overall, and 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) warn that increasing class sizes are particularly dangerous when coupled with current national trends towards increased student mobility, access to higher education, and internationalization of student composition. They point to impact of the 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 1,800 signatures on a petition criticizing the college's recent shift to a blended classroom model. Classes that tend to have between 800 and 2,000 students learn through a reduced class time format, which eliminates instructor-led lectures with the expectation that students spend more time learning with their peers outside of class to gain more thorough knowledge of the material (https://www.insidehighered.com/digital-learning/article/2018/09/21/blended-learning-model-university-central-florida-draws-business). From an institutional perspective, while the additional costs of smaller classes are viewed as prohibitively expensive as enrollment rises, results such as those presented here should not be ignored. Increased understanding of qualities that support learning and participation of students in small, medium, and large classes will improve effectiveness within institutional limitations.

Why do we observe gender differences in participation?

Our data show that the largest gender disparities in participation occur when instructors elicit 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 one second. With approximately one second, students must formulate a response and decide whether to participate, and many factors unrelated to content knowledge impact the decision to do so. Some of these factors may differentially affect men and women. For example, Cooper et al. (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 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., 2018; 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 perceived to be, negatively stereotyped (Steele et al., 2002). Extensive evidence from the precollege literature shows that regardless of how girls perform in a subject, they are more concerned about how instructors 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). This difference is apparent in several STEM disciplines at the college level, and likely plays a role in the observed skewed in-class participation towards 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, observers could not double check whether they categorized interactions correctly if they were unsure. An advantage of having observers in the classroom observing in real-time is a reduced uncertainty about student gender of participants, and observers could move if necessary to better identify students (which is not possible with a camera). While the person who trained all observers was the same (Ballen), we were only able to obtain reliability scores across observers at the University of Minnesota. Within the categories we used (voluntary response or response after group work) we consistently had very high inter-observer reliability at the University of Minnesota (>0.90), but this was not measured across all observers. Therefore we cannot rule out the possibility that reliability across other institutions was lower than at the University of Minnesota. However, for this reason we urge readers to find analyses of total responses the *most* reliable, which encompasses *all* types of interactions. Additionally, for responses where 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 than their ratio among those who volunteered. Instructors report that they did not knowingly do this, and results are similar between "individual spontaneous question" (i.e., in which a student asks an instructor an unprompted question or is 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 this study we only report male and female genders due to the limitations of our non-invasive observation methods, and we recognize we are unable to report more accurate gender identities. While 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 zero to four (i.e., likelihood of female participation was four times the probability of male participation), 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. While the outlier did not impact the overall results, it *created* a significant association between outcomes and whether students were in lower or upper division courses. Because we cannot completely rule out the possibility that the results which 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 (Supplementary material 2: Model selection summary, including outlier; Results tables, including outlier). While the current dataset 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’ from one second to between three to five seconds, Rowe (1974b) found that more students volunteer answers, and that students' answers were longer and more complex. Additionally, asking students to discuss questions in pairs or in groups lets students work through problems in a non-threatening environment, and practice expressing their opinions prior to being called upon (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 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 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 use a random number generator to spontaneously call on groups. Within student groups, randomly appointed ‘reporters’ can be responsible for voicing an answer on behalf of their group, which also takes responsibility off of the individual if the answer is incorrect (Cohen and Lotan, 2014). Instructors assign reporters based on 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 do 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 underlying factors that contribute to observed classroom disparities.

Conclusion

Our results align with previous work that calls for a halt on the continued expansion of large introductory ‘gateway’ courses in science (Achilles, 2012; Baker et al., 2016; Cuseo, 2007), and underscores the importance of continued empirical measurement of factors that either promote or counter equity in undergraduate STEM (Brewer & Smith, 2011; National Academies of Sciences and Medicine, 2016). In practice, the gender gap in participation means 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; Carter et al., 2017; Hinsley et al., 2017; Pritchard et al., 2014; Schmidt et al., 2017; Schmidt & Davenport, 2017), possibly contributing to a general tendency to undervalue the input of women in STEM (e.g., as grant recipients or speakers; Grunspan et al., 2016; Isbell et al., 2012).

Fortunately, while 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 females were more likely to participate after small group discussions and this effect was more pronounced when diverse teaching approaches were employed. Further, 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. Based on our results, large classes begin to negatively impact students when they are comprised of 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.

Acknowledgements. 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 from the National Science Foundation, RCN–UBE Incubator: Equity and Diversity in Undergraduate STEM; #1729935 awarded to S Cotner and CJ Ballen.

References

Achilles CM. 2012. Class-Size Policy: The STAR Experiment and Related Class-Size Studies. NCPEA Policy Brief. Volume 1, Number 2. *NCPEA Publications*

Ainsworth C. 2015. Sex redefined. *Nature* **518** (7539): 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): 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*

Ballen CJ, Danielsen M, Jørgensen C, Grytnes J-A, Cotner S. 2017a. Norway's gender gap: classroom participation in undergraduate introductory science. *Nordic Journal of STEM Education* **1** (1): 179–186

Ballen CJ, Lee D, Rakner L, Cotner S. 2018b. Politics a "Chilly" Environment for Undergraduate Women in Norway. *PS: Political Science & Politics*: 1–6

Ballen CJ, Wieman C, Salehi S, Searle JB, Zamudio KR. 2017b. Enhancing diversity in undergraduate science: Self-efficacy drives performance gains with active learning. *CBE-Life Sciences Education* **16** (4): ar56

Bates D, Mächler M, Bolker B, Walker S. 2014. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*

Beede DN, Julian TA, Langdon D, McKittrick G, Khan B, Doms ME. 2011. Women in STEM: A gender gap to innovation. *Economics and Statistics Administration Issue Brief* (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** (1): 2–39

Bettinger EP, Long BT. 2005. Do faculty serve as role models? The impact of instructor gender on female students. *American Economic Review* **95** (2): 152–157

Bornmann L, Mutz R, Daniel H-D. 2007. Gender differences in grant peer review: A meta-analysis. *Journal of Informetrics* **1** (3): 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, Washington, DC*
Brewer S, Klein JD. 2006. Type of positive interdependence and affiliation motive in an asynchronous, collaborative learning environment. *Educational Technology Research and Development* **54** (4): 331–354
Carter A, Croft A, Lukas D, Sandstrom G. 2017. Women's visibility in academic seminars: women ask fewer questions than men. *arXiv preprint arXiv:1711.10985*
Cohen EG, Lotan RA. 2014. *Designing Groupwork: Strategies for the Heterogeneous Classroom Third Edition*. 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** (2): 200–208
Cotner S, Ballen CJ. 2017. Can mixed assessment methods make biology classes more equitable? *PLoS ONE* **12** (12): 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** (5): 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. *The journal of higher education* **74** (1): 51–76
Cuseo J. 2007. The empirical case against large class size: adverse effects on the teaching, learning, and retention of first-year students. *The Journal of Faculty Development* **21** (1): 5–21
Dahlerup D. 1988. From a small to a large minority: women in Scandinavian politics. *Scandinavian Political Studies* **11** (4): 275–298
Dillon M, Kokkelenberg EC, Christy SM. 2002. The Effects of Class Size on Student Achievement in Higher Education: Applying an Earnings Function.
Eddy SL, Brownell SE, Wenderoth MP. 2014. Gender gaps in achievement and participation in multiple introductory biology classrooms. *CBE-Life Sciences Education* **13** (3): 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** (23): 8410–8415

Freeman S, Haak D, Wenderoth MP. 2011. Increased course structure improves performance in introductory biology. *CBE Life Sci Educ* **10** (2): 175–186 DOI: 10.1187/cbe.10-08-0105

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** (2): 1–16 DOI: 10.1371/journal. pone.0148405

Haak DC, HilleRisLambers J, Pitre E, Freeman S. 2011. Increased structure and active learning reduce the achievement gap in introductory biology. *Science* **332** (6034): 1213–1216 DOI: 10.1126/science.1204820

Hallgren KA. 2012. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology* **8** (1): 23

Hinsley A, Sutherland WJ, Johnston A. 2017. Men ask more questions than women at a scientific conference. *PloS one* **12** (10): e0185534

Ho DE, Kelman MG. 2014. Does class size affect the gender gap? a natural experiment in law. *The Journal of Legal Studies* **43** (2): 291–321

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

Isbell LA, Young TP, Harcourt AH. 2012. Stag parties linger: continued gender bias in a female-rich scientific discipline. *PLoS One* **7** (11): 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** (1): 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** (2): 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** (3): 332–342

Kuh GD, Hu S. 2001. The effects of student-faculty interaction in the 1990s. *The review of higher education* **24** (3): 309–332

Ledin A, Bornmann L, Gannon F, Wallon G. 2007. A persistent problem: Traditional gender roles hold back female scientists. *EMBO reports* **8** (11): 982–987

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

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** (3)

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** (2): 396

Premo J, Cavagnetto A. 2018. Priming students for whole-class interaction: Using interdependence to support behavioral engagement. *Social Psychology of Education*: 1–21

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

Rowe MB. 1974a. Wait-time and rewards as instructional variables, their influence on language, logic, and fate control: Part one-wait-time. *Journal of research in science teaching* **11** (2): 81–94

Rowe MB. 1974b. Relation of wait-time and rewards to the development of language, logic, and fate control: Part II-Rewards. *Journal of research in science teaching* **11** (4): 291–308

Rusbult CE, Van Lange PAM. 2008. Why we need interdependence theory. *Social and Personality Psychology Compass* **2** (5): 2049–2070

Schanzenbach DW. 2014. Does Class Size Matter?

Schmidt SJ, Davenport JRA. 2017. Who asks questions at astronomy meetings? *NATURE* **1** (0153): 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 preprint arXiv:1704.05260*

Simpson EH. 1949. Measurement of diversity. *nature*

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** (5910): 122 LP-124 Available at: 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. In *Advances in Experimental Social Psychology*Academic Press; 379–440.

Tanner KD. 2013. Structure matters: twenty-one teaching strategies to promote student engagement and cultivate classroom equity. *CBE-Life Sciences Education* **12** (3): 322–331

Team RC. 2013. R: A language and environment for statistical computing

Theobald E. 2018. Students are rarely independent: When, why, and how to use random effects in discipline-based education research. *CBE—Life Sciences Education* **17** (3): rm2

Wold A, WENNERÁS C. 2010. Nepotism and sexism in peer-review. In *Women, Science, and Technology*Routledge; 64–70.

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