Do Small Classes in Higher Education Reduce Performance Gaps in STEM?

CISSY J. BALLEN, STEPFANIE M. AGUILLON, REBECCA BRUNELLI, ABBY GRACE DRAKE, DEENA WASSENBERG, STACEY L. WEISS, KELLY R. ZAMUDIO, AND SEHOYA COTNER

Performance gaps in science are well documented, and an examination of underlying mechanisms that lead to underperformance and attrition of women and underrepresented minorities (URM) may offer highly targeted means to promote such students. Determining factors that influence academic performance may provide a basis for improved pedagogy and policy development at the university level. We examined the impact of class size on students in 17 biology courses at four universities. Although the female students underperformed on high-stakes exams compared with the men as class size increased, the women received higher scores than the men on nonexam assessments. The URM students underperformed across grade measures compared with the majority students regardless of class size, suggesting that other characteristics of the education environment affect learning. Student enrollment is expected to increase precipitously in the next decade, underscoring the need to prioritize individual student potential rather than yield to budget constraints when considering equitable pedagogy and caps on classroom sizes.

Keywords: education, assessments, behavioral science

Universities face the unique challenge of educating students from increasingly diverse backgrounds who may excel in different educational contexts. Recent efforts to better serve diverse classrooms include changes in instruction such as active learning (Haak et al. 2011, Ballen et al. 2017a) and course-based undergraduate research experiences (Lopatto 2007, Ballen et al. 2017c). To provide effective instructional practices for all, we must continue to identify practical steps to promote the success of qualified students from historically underserved demographics in science, technology, engineering, and mathematics (STEM), such as women and underrepresented minority students (African American, Hispanic, Native American, or Pacific Islander; hereafter "URM").

If our goal is to achieve diversity in STEM, coursework should ideally nurture individual potential rather than "weed out" less prepared students at the start of an undergraduate degree (Suresh 2006, Mervis 2011, Koester et al. 2016). Using 16 years of data from a liberal arts college, Rask and Tiefenthaler (2008) demonstrated that the students' grades influenced their decision to continue within their major. Although lower grades led to lower persistence for all students, the female students with low grades were more likely than the males to abandon the discipline and pursue a different major. A second longitudinal study showed that negative experiences in introductory science courses were cited as the primary reason for declining interests in obtaining a science degree among the women and URM students (Barr et al. 2008). Women and URM students also face other well-documented challenges unrelated to academic competency, such as discrimination (Steele J et al. 2002, Moss-Racusin et al. 2012, Milkman et al. 2015, Grunspan et al. 2016), feelings of exclusion (Hall and Sandler 1982, Hurtado and Ruiz 2012), imposter syndrome (Clance 1985), test anxiety (Ballen et al. 2017b), and stereotype threat (Steele CM and Aronson 1995, Steele CM 1997, Schmader 2002). All of these contribute to the well-documented higher attrition rates of women and URM students across STEM disciplines (May and Chubin 2003, Alexander et al. 2009, Beede et al. 2011, Eddy et al. 2014, Ballen and Mason 2017) and university campuses (Smith 2000, Anderson and Kim 2006, Griffith 2010, Olson and Riordan 2012). Education research has also identified examples of learning contexts that counteract the psychosocial barriers faced disproportionately by women and URM students, including opportunities to interact with role models in and out of the classroom (Fried and MacCleave 2009, Stout et al. 2011), interventions in social belonging (Walton et al. 2015), peer mentoring (Snyder and Wiles 2015), and for females, schools with higher percentages of female STEM graduate students (Griffith 2010). Therefore, it is essential

BioScience XX: 1–8.
doi:10.1093/biosci/biy056

that we identify obstacles that specifically affect underrepresented students as a means of finding interventions that promote all students' success in STEM.

Class size, an often overlooked variable, is worthy of careful consideration because previous research suggests it influences student performance (Glass 1982, Kokkelenberg et al. 2008, Ho and Kelman 2014) and, unlike other variables, is subject to legislative action. At least 24 states have mandated or incentivized class-size reduction in American K–12 classrooms (Whitehurst and Chingos 2011). At the undergraduate level, universities are constantly faced with decisions on how to allocate faculty time to best serve their undergraduate population. Recent changes in course content delivery—such as the rise of online classes (e.g., massive open online courses, or MOOCs) and hybrid online courses—are the direct result of an increased demand for access to education (Kena et al. 2016). The imminent growth in enrollment at degree-granting institutions (Kena et al. 2016) underscores the urgent need to quantify the effects of class sizes on undergraduate students. Here, using data from 17 biology courses at four institutions, we examine the extent that class size affects achievement gaps for female and URM undergraduates.

We address three questions by focusing on performance gaps between male and female students and between URM and majority students: (1) Does class size influence performance on exams? 2) Does class size influence performance on nonexam methods of assessment? (3) Does class size influence final course grade?

Data collection

Administrative data were obtained from 17 lower-division biology courses taken by 1836 students in fall 2016 (minimum class size n = 40, maximum n = 239; figure 1). To establish a collaborative research group, we solicited participation through an existing professional network from biology instructors who teach majors or nonmajors from a diverse range of institutions, and we received data from California State University, Chico; Cornell University; the University of Minnesota, Twin Cities; and the University of Puget Sound. The network was sustained through a Research Coordination Network funded by the National Science Foundation (RCN–UBE Incubator: Equity and Diversity in Undergraduate STEM; award #1729935). We compared (a) pooled exam grades, (b) pooled assessments of student knowledge other than exams (hereafter "nonexam grades"; e.g., discussion sections, laboratories, online activities, written assignments, low-stakes quizzes, as well as active-learning, in-class activities), and (c) final course grades, which reflect cumulative performance in all aspects of the course. We present analyses with transformed z-scores (a measure of how many standard deviations a value is from the class section's mean score) for ease of interpretation.

Statistical analyses: Linear mixed-effects models

We used linear mixed-effects models to compare exam performance, performance on nonexam assessments, and total course performance across the four universities. The data in this study are hierarchically nested because a student's exam performance is likely to be more similar to a classmate's performance than to that of a student outside of their class, because students in the same class share the same assessments (Kreft et al. 1998). Similarly, students in biology classes at one university may perform or be assessed in the same way as compared to students in biology classes at another university. For this reason, we use multilevel modeling to account for the nonindependence of data in nested-data structures (Paterson and Goldstein 1991, Kreft et al. 1998).

Akaike's information criterion (AIC) was used to determine model fit in a multimodel inference technique. AIC estimates the goodness of fit of each model given our sample (Akaike 1974) and allows us to rank models on the basis of this estimation using AIC differences ($\Delta i = \mathrm{AIC}_{\mathrm{model}}\, i$ – minAIC, where minAIC is the model with the smallest AIC value). Models with a $\Delta i > 10$ are considered poor predictors compared with the best model, so we only present results with small Δi values for brevity (table 1). We were interested in the interaction of class size with gender (SGender, a factor with two levels) and with URM status (a factor with two levels). Therefore, our model initially included those three main effects (SGender, URM status, and class size) and two interaction effects (SGender*class size and URM status*class size).

In addition, we tested whether the following variables improved the fit of the model for the given set of data: (a) an interaction between student gender identity and URM status (SGender*URM status); (b) instructor gender identity (IGender, a factor with three levels including female, male, or multiple instructor genders—in other words, more than one instructor for the course in question who did not identify as the same gender); (c) an interaction between student gender identity and instructor gender identity (SGender*IGender); (d) an interaction between student gender identity, URM status, and class size (SGender*URM status*class size); and (e) age. Only students with a complete set of these variables were included in these analyses. All models included random effects for university, class ID (nested within university), and instructor ID (nested within classes and university). Random effects were tested for significance by removing one random factor at a time and taking the difference between the –2 log likelihoods. This was tested against a chi-square distribution with one degree of freedom (per removed random factor). Instructor ID was removed from the analysis as a random effect.

We explored all possible models and chose the most parsimonious model that best fit the data in accordance to AIC model-selection statistics (table 1). The AIC estimates indicated that the elimination of the URM*class size interaction resulted in better fit models, and so the interaction was backward eliminated from the final models ($p > .25$; see results). We used Bonferroni corrected post hoc pairwise comparisons to clarify the performance outcomes of the

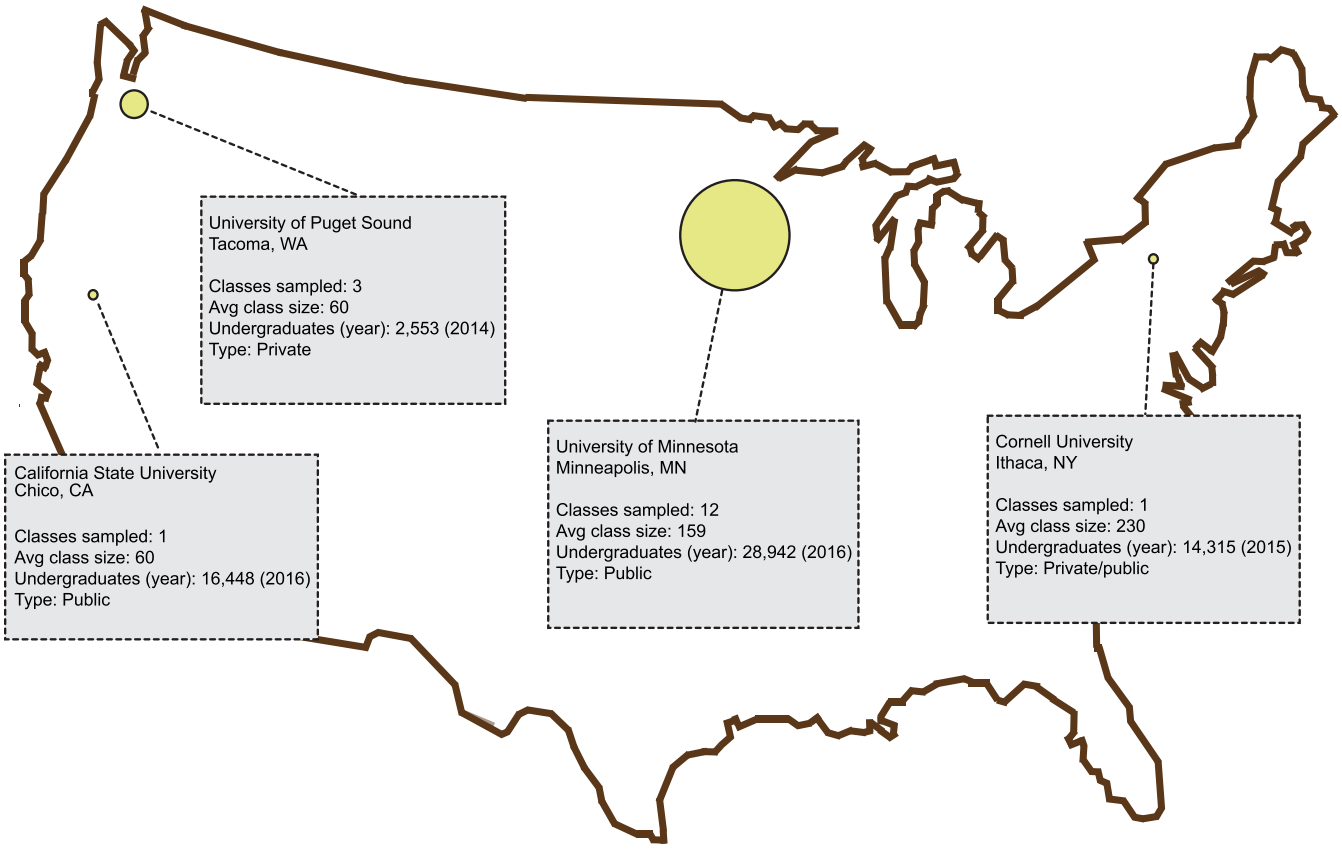

Figure 1. Four universities participated in the current study, representing diverse geographic locations across the United States. The circle sizes are proportional to the number of classes sampled from each institution.

students based on gender and URM status. We performed all statistical analyses using SPSS software version 24 (SPSS Inc., Chicago, Illinois).

Results

We used mixed-model analyses to compare the students' combined exam grade, nonexam grade, and total course grade in the fall 2016 semester (figure 2; supplemental table S1–S3). First, we observed a nonsignificant interaction effect of URM status and class size on metrics of performance.

When we removed the interaction from the models, URM status became a significant predictor of performance (combined exam grade $B = 0.417$, $t(1377) = 6.01$, $p < .001$, $SE = 0.069$; nonexam grade $B = 0.262$, $t(1533) = 3.83$, $p < .001$, $SE = 0.069$; final course score $B = 0.407$, $t(1522) = 5.87$, $p < .001$, $SE = 0.069$). These results suggest that the URM students' exam scores on average were 0.42 standard deviation lower than those of non-URM students, and their nonexam scores were on average 0.26 standard deviation lower than those of non-URM students. Bonferroni corrected post hoc pairwise comparisons, presented from the final models, show the URM students underperforming on all performance metrics compared with the non-URM students (table 2; hereafter, "underperform" is used to describe raw gaps, and not those for which some measure of student academic ability or preparation is controlled). Second, we observed a significant interaction between gender and class size, such that as class size increased, the women underperformed on exams (SGender*class size $B = -0.145$, $t(1599) = -2.89$, $p = .004$, $SE = 0.050$; figure 2 inset) and in the course overall ($B = -0.108$, $t(1649) = -2.16$, $p = .031$, $SE = 0.050$) compared with men. We also found that the women obtained higher nonexam grades ($B = 0.217$, $t(1731) = 4.60$, $p < .001$, $SE = 0.047$) compared with men, regardless of class size.

Next, we explored whether women are underperforming on exams because those tests are higher stakes in larger classes—that is, they account for a larger proportion of the grade. To investigate this, we examined the correlation between class size and the percentage of the students' final course grades that were from their performance on exams. We did not find a strong correlation (Pearson correlation = –0.386; $p = .126$). This result runs counter to what one would expect because of the courses included in this sample and is probably not representative of most lower-division lecture courses, in which exams generally account for a larger proportion of final course grades (Koester et al. 2016). Finally, to test whether our results are the same within one institution, we isolated 12 lower-division classes from the University of Minnesota that varied in class size. In these classes, all exams had identical multiple-choice formats. We

Table 1. Best models for predicting performance metrics across four universities using AIC model selection.

Rank	Model: Combined exam grades	AIC	Δ*i*	Relative likelihoods	w_i
1	URM status + class size + SGender + class size*Sgender	4885.468	0.000	1.000	0.935
2	URM status + class size + Sgender + class size*Sgender + age	4891.961	6.493	0.039	0.036
3	URM status + class size + Sgender + class size*Sgender + Sgender*URM status + age	4892.347	6.879	0.032	0.030
	Model: Non-exam grades				
1	URM status + class size + SGender + class size*Sgender	4835.231	0.000	1.000	0.885
2	URM status + class size + SGender + class size*Sgender age	4839.840	4.609	0.100	0.088
3	URM status + class size + SGender + class size*Sgender + class size* URM status + age	4842.562	7.331	0.026	0.023
	Model: Final course grade				
1	URM status + SGender	4826.220	0.000	1.000	0.926
2	URM status + class size + SGender	4831.668	5.448	0.066	0.061
3	URM status + class size + Sgender + age	4836.220	10.000	0.007	0.006
4	Sgender	4837.260	11.040	0.004	0.004
5	URM status + class size + SGender + class size*Sgender + class size* URM status	4837.736	11.516	0.003	0.003

Note: Compared with the first model, models with an $\Delta_i > 10$ are considered poor predictors, so we do not report them here. The Akaike weights, w_i, represent probabilities that a given model is the best model under repeated sampling.

found the same main results across assessment types within one institution as we observed across all institutions (supplemental tables S4–S6). Therefore, as was the case across universities, increasing class size was negatively correlated with female performance, and URM status significantly predicted performance outcomes within our most sampled university.

One possibility is that the positive effects we observe from the students in small classes is due to increased active learning and student interactions with the instructor in smaller classes, which may influence student performance (e.g., Haak et al. 2014, Ballen et al. 2017c). Using data collected for 9 of the 17 courses (supplemental table S7), we used a linear regression to examine the relationship between class size and the total number of student–instructor interactions per class period. Results from the linear regression were not conclusive. First, when we included all of the schools in our analysis, we found a significant relationship between the two variables (supplemental figure S1; Pearson correlation = –0.72; $p = .028$), such that the students interacted more with their instructors in smaller classes. However, when we isolated classes within the University of Minnesota, the correlation was no longer significant (Pearson correlation = 0.24; $p = .645$). Class size likely influences the frequency with which students interact with their instructors, and this may be why small class sizes appear to disproportionately benefit women in our sample. Future work will profit from a thorough examination of the relationship between class size, active learning, and performance gaps.

Conclusions

We compared female and male exam performance, nonexam performance, and total course performance across four universities and found that as class sizes increased, the women underperformed on exams and final course grades compared with the men in their classes. However, the female students outperformed the males regardless of class size on nonexam scores that contributed to total course grades. We did not find a similar effect of class size on students based on minority status. Across class size and assessment type, the URM students underperformed relative to the non-URM students (table 2).

Reasons for the pervasive disparity between URM and non-URM students are likely complex and multifaceted but may include differences in incoming academic preparation (Ballen and Mason 2017), economic hardship (Cabrera et al. 1992), university campus social climate (Gloria et al. 1999), and low representation in the classroom or discipline (Braxton et al. 2011). The underrepresentation of URM individuals in the STEM workforce (Landivar 2013) underscores the urgent need for effective approaches that promote students who are racial or ethnic minorities (Brewer and Smith 2011).

Although our findings do not suggest tractable solutions to racial disparities in STEM, they do suggest strategies for mitigating gender biases. Specifically, to increase female retention in STEM, we recommend offering smaller classes and emphasizing nonexam points, especially in lower-division classes that serve as gateway courses to students' major fields of study. In these gateway courses, students are often weeded out because students' perceived or actual academic performance suffers in those environments (Baker et al. 2016).

A review by Cuseo (2007) identified five reasons that large classes have adverse effects on some students: (1) fewer

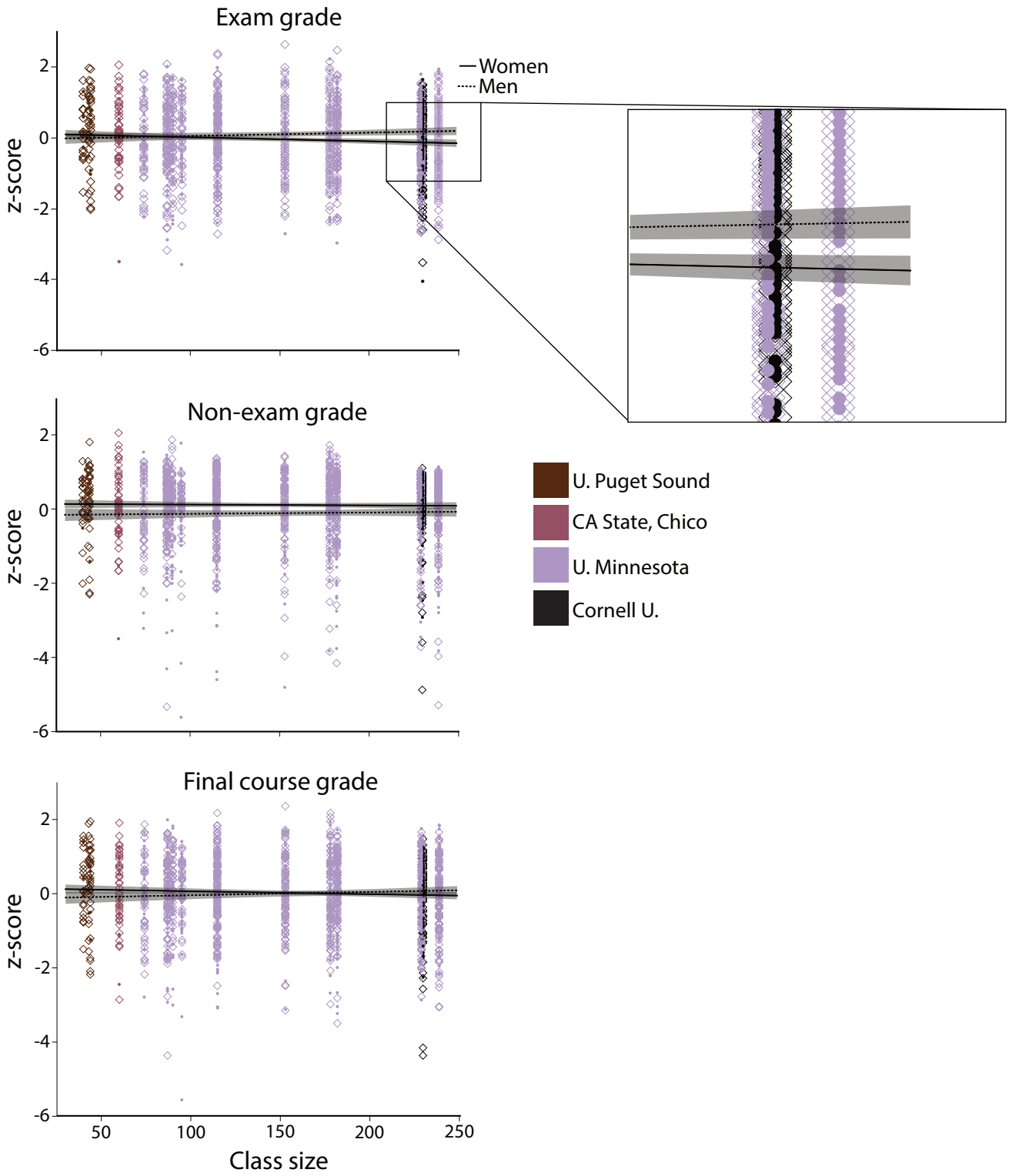

Figure 2. The effects of class size on exam grade z-scores, nonexam grade z-scores, and final course grade z-scores for women (solid line) and men (dashed line). The colors represent different universities: the University of Puget Sound (brown); California State University, Chico (pink); the University of Minnesota, Twin Cities (purple); and Cornell University (black).

opportunities for students to interact with course material, (2) fewer opportunities for students to interact with the instructor, (3) reduced opportunities for instructors to challenge students, (4) lower overall student satisfaction with the learning experience, and (5) lower satisfaction with the instructor according to student evaluations (Cuseo 2007). Future research will benefit from a close examination of the consequences of these factors and whether they

Table 2. A least-squares means comparison of the relative performance of students who differ on the basis of their racial minority status (underrepresented minority, URM, or non-URM) in different class sizes (50 students, 150 students, or 250 students).

	URM						non-URM					
	Mean (M)	Standard error (SE)	M	SE	M	SE	M	SE	M	SE	M	SE
Class size	**50**		**150**		**250**		**50**		**150**		**250**	
Combined exam grade	–0.285	0.08	–0.295	0.07	–0.305	0.08	0.140	0.06	0.130	0.05	0.120	0.06
Nonexam grade	–0.165	0.08	–0.165	0.08	–0.166	0.09	0.086	0.07	0.086	0.06	0.085	0.07
Total course grade	–0.262	0.09	–0.264	0.08	–0.266	0.09	0.132	0.07	0.130	0.06	0.128	0.07
n			261						1575			

Note: Measures are standardized and reflect performance relative to the mean of the class; the positive scores are students who overperformed in standard deviations from the mean, and the negative scores represent those who underperformed relative to the mean. Our data indicate that URM students underperform across all metrics compared with non-URM students, but unlike female students, their performance is not affected by class size, suggesting that factors other than class size negatively influence URM student performance.

respond to experimental class-size manipulations. We do recognize the reality of budgetary constraints and the fact that larger classes are often the simplest solution to fiscal crises. However, when large classes are a "necessary evil," instructors can minimize the negative consequences of large classes via evidence-based interventions. For example, in large lecture settings, students can have more opportunities to interact with lecture material and the instructor via numerous instant-feedback strategies (e.g., the immediate feedback assessment technique, Cotner et al 2008a; classroom response systems, Cotner et al 2008b, Lewin et al 2016, Knight et al 2016; and plicker cards, Howell et al 2017) and low-stakes—or no-stakes—formative assessments (e.g., 1-minute papers, worksheets, and concept maps; Angelo and Cross 1993).

Because in our data set, the female students excelled at nonexam assessments of the course material regardless of class size, an alternative strategy to promote women in STEM may be to make nonexam scores a larger component of the final course grade (Koester et al. 2016). Recent work shows that traditional exams do not accurately capture student mastery of the cognitive skills required to do science and that they exacerbate existing gaps in performance (Stanger-Hall 2012, Moneta-Koehler et al. 2017). Furthermore, women are adversely affected by test anxiety, which in itself is higher in women than in their male counterparts (Ballen et al. 2017b). Therefore, if our aim is to reward ongoing preparation and cooperative group work rather than performance on a few high-stakes exams, these assignments will nurture those qualities and work habits in developing scientists. For instructors who teach large classes, the challenge will be to develop scalable assignments that can effectively evaluate students' learning. Despite these challenges, our data show that an effective way for instructors to reduce gender gaps in their classrooms is to experiment with strategies to tailor the learning environment to their student population.

Research demonstrating the negative impacts of large classes on students reinforces conceptual arguments against these classes (Glass and Smith 1979, Glass 1982, Achilles 2012, Ho and Kelman 2014, Schanzenbach 2014, Baker et al. 2016), and can inform policy related to education. The state of Minnesota, in which the majority of classes were sampled, has historically taken innovative approaches to improving its schools (Mazzoni 1993). In fact, the state's former governor, Jesse "The Body" Ventura, campaigned on an education platform that declared "the best way to solve most of our educational problems is to reduce class size" (Ventura 2000). Nationally, schools aim to keep class sizes low, but according to the National Center for Education Statistics, total enrollment at public and private degree-granting postsecondary institutions is expected to increase 15% between 2014 and 2025 (Kena et al. 2016). Although it may be tempting to increase the number of students per class section in order to decrease costs, the consequences on student learning and performance must be carefully considered. Note that our classes range in size from 40 to over 200 students. Therefore, a class of 50–100 students is associated, in our model, with more equitable performance than is one with 200 or more students; in other words, a "smaller" class is likely still cost-effective. Future work will conduct similar investigations into the effects of class size on students of low socioeconomic status and first-generation college students.

This work has limitations that warrant consideration. First, we were unable to control for incoming student preparation (e.g., precourse measures such as the SAT or cumulative GPA) for all students across universities. Previous work finds that incoming preparation predicts performance and retention across institutions (Bonous-Hammarth 2000, Ballen and Mason 2017, Ballen et al. 2017b, Easton et al. 2017). However, by normalizing performance across cohorts, we show the achievement gaps in course grades as they are corrected in magnitude. Second, to test the generality of these results, it will be important to test a wider range

of universities nationally and internationally. Although our data set is subject to some biases, these collaborative efforts among universities allow for much larger data sets—across a broad sample of university types—that would not be possible within one institution. Thus, multi-institution efforts allow for meaningful comparisons and have considerable potential to illuminate the nature of persistent demographic gaps within classrooms, as well as gaps in institutional representation in the STEM workforce. Finally, many other variables may contribute to student performance that we did not include in our analysis, including teaching strategy (e.g., active or traditional lecturing; Haak et al. 2011), classroom social climate (Crawford and MacLeod 1990, Grunspan et al. 2016), campus social climate (Hall and Sandler 1984), and opportunity for academic support outside of the classroom (e.g., tutorials or peer mentoring; Snyder et al. 2016). Future work will also benefit from a focus on the underlying mechanisms that explain the observed gender gaps in large classes at the undergraduate level.

Despite these limitations, we detect an interaction effect between gender and class size, such that women are negatively affected by large class sizes in ways that men are not. These findings add an equity dimension to previous work citing the benefits of smaller classes. This aspect of smaller-class impacts may be especially compelling to administrators, curriculum committees, or legislators who are motivated to eliminate the gender gaps in performance that plague higher education.

Acknowledgments

We thank Daniel Baltz for help with data organization and interpretation; J. D. Walker and Lauren Sullivan for statistical support; and Gregor Siegmund, Paula Soneral, Brian Wisenden, Daniel Stovall, Michelle Mabry, Steven Karafit, Denise Monti, Danielle Grunzke, Leslie Saucedo, Gregory Johnson, Jorge Tomasevic, and Heather Patterson for help with student data. We received IRB exemption to work with student data at all universities (IRB protocol UMN, no. 1405E50826; CU, no. 1410005010; UPS, no. 1617-006; CSU, no. 461).

Funding statement

Coordination and data management were supported by funding to C.J.B. and S.C. from the National Science Foundation (NSF) Research Coordination Network (RCN-UBE Incubator: Equity and Diversity in Undergraduate STEM; Award #1729935).

Supplemental material

Supplementary 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 (NCPEA) Policy Brief, vol. 1. NCPEA.

Akaike H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19: 716–723.

Alexander C, Chen E, Grumbach K. 2009. How leaky is the health career pipeline? Minority student achievement in college gateway courses. Academic Medicine 84: 797–802.

Anderson E, Kim D. 2006. Increasing the Success of Minority Students in Science and Technology. American Council on Education.

Angelo TA, Cross KP. 1993. Classroom Assessment Techniques: A Handbook for Faculty. National Center for Research to Improve Postsecondary Teaching and Learning.

Baker BD, Farrie D, Sciarra DG. 2016. Mind the Gap: 20 Years of Progress and Retrenchment in School Funding and Achievement Gaps. Educational Testing Service.

Ballen CJ, Mason NA. 2017. Longitudinal analysis of a diversity support program in biology: A national call for further assessment. BioScience 67: 367–373.

Ballen CJ, Blum JE, Brownell S, Hebert S, Hewlett J, Klein JR, McDonald EA, Monti DL, Nold SC, Slemmons KE. 2017a. A call to develop course-based undergraduate research experiences (CUREs) for nonmajors courses. CBE–Life Sciences Education 16 (art. mr2).

Ballen CJ, Salehi S, Cotner S. 2017b. Exams disadvantage women in introductory biology. PLOS ONE 12 (art. e0186419).

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

Barr DA, Gonzalez ME, Wanat SF. 2008. The leaky pipeline: Factors associated with early decline in interest in premedical studies among underrepresented minority undergraduate students. Academic Medicine 83: 503–511.

Beede D, Julian T, Langdon D, McKittrick G, Khan B, Doms M. 2011. Women in STEM: A Gender Gap to Innovation. US Department of Commerce. Economics and Statistics Administration Issue Brief no. 04-11.

Bonous-Hammarth M. 2000. Pathways to success: Affirming opportunities for science, mathematics, and engineering majors. Journal of Negro Education 69: 92–111.

Braxton JM, Hirschy AS, McClendon SA. 2011. Understanding and Reducing College Student Departure. Wiley. ASHE–ERIC Higher Education Report, vol. 30.

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

Cabrera AF, Nora A, Castaneda MB. 1992. The role of finances in the persistence process: A structural model. Research in Higher Education 33: 571–593.

Clance PR. 1985. The Impostor Phenomenon: Overcoming the Fear That Haunts Your Success. Peachtree.

Cotner S, Baepler P, Kellerman A. 2008a. Scratch this! The IF-AT as a technique for stimulating group discussion and exposing misconceptions. Journal of College Science Teaching 37: 48–53.

Cotner SH, Fall BA, Wick SM, Walker JD, Baepler PM. 2008b. Rapid feedback assessment methods: Can we improve engagement and preparation for exams in large-enrollment courses? Journal of Science Education and Technology 17: 437–443. doi:10.1007/s10956-008-9112-8

Crawford M, MacLeod M. 1990. Gender in the college classroom: An assessment of the "chilly climate" for women. Sex Roles 23: 101–122.

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.

Easton JQ, Johnson E, Sartain L. 2017. The predictive power of ninth-grade GPA. University of Chicago Consortium on School Research.

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.

Fried T, MacCleave A. 2009. Influence of role models and mentors on female graduate students' choice of science as a career. Alberta Journal of Educational Research 55: 482–496.

Glass GV. 1982. School Class Size: Research and Policy. Sage.

Glass GV, Smith ML. 1979. Meta-analysis of research on class size and achievement. Educational Evaluation and Policy Analysis 1: 2–16.

Gloria AM, Kurpius SER, Hamilton KD, Willson MS. 1999. African American students' persistence at a predominantly white university: Influences of social support, university comfort, and self-beliefs. Journal of College Student Development 40: 257–268.

Griffith AL. 2010. Persistence of women and minorities in STEM field majors: Is it the school that matters? Economics of Education Review 29: 911–922.

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: 1–16.

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.

Hall RM, Sandler BR. 1982. The Classroom Climate: A Chilly One for Women? Association of American Colleges, Project on the Status and Education of Women.

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

Hurtado S, Ruiz A. 2012. The climate for underrepresented groups and diversity on campus. American Academy of Political and Social Science 634: 190–206.

Kena G, et al. 2016. The Condition of Education. National Center for Educational Statistics.

Knight JK, Wise SB, Sieke S. 2016. Group random call can positively affect student in-class clicker discussions. CBE–Life Sciences Education 15 (art. ar56). doi:10.1187/cbe.16-02-0109

Koester BP, Grom G, and McKay TA. 2016. Patterns of Gendered Performance Difference in Introductory STEM Courses. Cornell University Library. (17 May 2018; h*ttps://arxiv.org/abs/1608.07565*)

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.

Kreft I, de Leeuw J. 1998. Introducing Multilevel Modeling. Sage.

Landivar LC. 2013. Disparities in STEM employment by sex, race, and Hispanic origin. Education Review 29: 911–922.

Lewin JD, Vinson EL, Stetzer MR, Smith MK. 2016. A Campus-Wide Investigation of Clicker Implementation: The Status of Peer Discussion in STEM Classes. CBE–Life Sciences Education 15 (art. ar6). doi:10.1187/cbe.15-10-0224

Lopatto D. 2007. Undergraduate research experiences support science career decisions and active learning. CBE–Life Sciences Education 6: 297–306.

May GS, Chubin DE. 2003. A retrospective on undergraduate engineering success for underrepresented minority students. Journal of Engineering Education 92: 27–39.

Mazzoni TL. 1993. The changing politics of state education policy making: A 20-year Minnesota perspective. Educational Evaluation and Policy Analysis 15: 357–379.

Mervis J. 2011. Weed-out courses hamper diversity. Science 334: 1333–1333.

Milkman KL, Akinola M, Chugh D. 2015. What happens before? A field experiment exploring how pay and representation differentially shape bias on the pathway into organizations. Journal of Applied Psychology 100: 1678–1712.

Moneta-Koehler L, Brown AM, Petrie KA, Evans BJ, Chalkley R. 2017. The limitations of the GRE in predicting success in biomedical graduate school. PLOS ONE 12 (art. e0166742).

Moss-Racusin CA, Dovidio JF, Brescoll VL, Graham MJ, Handelsman J. 2012. Science faculty's subtle gender biases favor male students. Proceedings of the National Academy of Sciences 109: 16474–16479.

Olson S, Riordan DG. 2012. Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics. Executive Office of the President.

Paterson L, Goldstein H. 1991. New statistical methods for analysing social structures: An introduction to multilevel models. British Educational Research Journal 17: 387–393.

Rask K, Tiefenthaler J. 2008. The role of grade sensitivity in explaining the gender imbalance in undergraduate economics. Economics of Education Review 27: 676–687.

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

Schmader T. 2002. Gender identification moderates stereotype threat effects on women's math performance. Journal of Experimental Social Psychology 38: 194–201.

Smith TY. 2000. 1999–2000 SMET Retention Report: The Retention and Graduation Rates of 1992–98 Entering Science, Mathematics, Engineering and Technology Majors in 119 Colleges and Universities. University of Oklahoma Center for Institutional Data Exchange and Analysis.

Snyder JJ, Wiles JR. 2015. Peer led team learning in introductory biology: Effects on peer leader critical thinking skills. PLOS ONE 10 (art. e0115084).

Snyder JJ, Sloane JD, Dunk RD, Wiles JR. 2016. Peer-led team learning helps minority students succeed. PLOS Biology 14 (art. e1002398).

Stanger-Hall KF. 2012. Multiple-choice exams: An obstacle for higher-level thinking in introductory science classes. CBE–Life Sciences Education 11:294–306.

Steele CM. 1997. A threat in the air: How stereotypes shape intellectual identity and performance. American Psychology 52: 613–629.

Steele CM, Aronson J. 1995. Stereotype threat and the intellectual test performance of African Americans. Journal of Personality and Social Psychology 69: 797–811.

Steele J, James JB, Barnett RC. 2002. Learning in a man's world: Examining the perceptions of undergraduate women in male-dominated academic areas. Psychology of Women Quarterly 26: 46–50.

Stout JG, Dasgupta N, Hunsinger M, McManus MA. 2011. STEMing the tide: Using ingroup experts to inoculate women's self-concept in science, technology, engineering, and mathematics (STEM). Journal of Personality and Social Psychology 100: 255–270.

Suresh R. 2006. The relationship between barrier courses and persistence in engineering. Journal of College Student Retention: Research, Theory and Practice 8: 215–239.

Ventura J. 2000. I Ain't Got Time to Bleed: Reworking the Body Politic from the Bottom Up. Villard.

Walton GM, Logel C, Peach JM, Spencer SJ, Zanna MP. 2015. Two brief interventions to mitigate a "chilly climate" transform women's experience, relationships, and achievement in engineering. Journal of Educational Psychology 107: 468–485.

Whitehurst GJ, Chingos MM. 2011. Class Size: What Research Says and What It Means for State Policy. Brookings Institution.

Cissy J. Ballen (balle027@umn.edu), Deena Wassenberg, and Sehoya Cotner are affiliated with Department of Biology Teaching and Learning at the University of Minnesota, in Minneapolis. 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. SMA is also with the Cornell Lab of Ornithology. Rebecca Brunelli is with the Department of Biological Sciences at California State University, in Chico. Stacey L. Weiss is affiliated with the Department of Biology at the University of Puget Sound, in Tacoma, Washington.