Do small classes in higher education reduce performance gaps in STEM?

Cissy J. Ballen1,2, Stephanie M. Aguillon3,4, Rebecca Brunelli5, Abby Grace Drake3, Deena Wassenberg1, Stacey L. Weiss6, Kelly R. Zamudio3, and Sehoya Cotner1

1 Department of Biology Teaching and Learning, University of Minnesota, Minneapolis, MN
2 balle027@umn.edu
3 Department of Ecology & Evolutionary Biology, Cornell University, Ithaca, NY
4 Cornell Lab of Ornithology, Ithaca, NY
5 Department of Biological Sciences, California State University, Chico, CA
6 Department of Biology, University of Puget Sound, Tacoma, WA

Manuscript published as an Article in Bioscience

Word count: 5,511
Abstract

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. While female students underperformed on high-stakes exams compared to men as class size increased, women received higher scores than men on non-exam assessments. URM students underperformed across grade measures compared to 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.
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 (Ballen et al. 2017a; Haak et al. 2011) and course-based undergraduate research experiences (Ballen et al. 2017b, Lopatto 2007). 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 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 (Koester et al. 2016, Mervis 2011, Suresh 2006). Using 16 years of data from a liberal arts college, Rask and Tiefenthaler (2008) demonstrated that students’ grades influenced their decision to continue within their major. While lower grades led to lower persistence for all students, female students with low grades were more likely to abandon the discipline and pursue a different major than males. 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 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 (Grunspan et al. 2016, Milkman et al. 2015, Moss-Racusin et al. 2012, Steele Jennifer et al. 2002), feelings of exclusion (Hall and Sandler 1982, Hurtado and Ruiz 2012), imposter syndrome (Clance 1985), test anxiety (Ballen et al., 2017) and stereotype threat (Schmader 2002, Steele 1997, Steele and Aronson 1995). All of these contribute to the well
documented higher attrition rates of women and URM students across STEM disciplines (Alexander et al. 2009, Ballen and Mason 2017, Beede et al. 2011, Eddy et al. 2014, May and Chubin 2003) and university campuses (Anderson and Kim 2006, Griffith 2010, Olson and Riordan 2012, Smith 2000). 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). Thus, it is essential 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, Ho and Kelman 2014, Kokkelenberg et al. 2008) 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 – e.g., the rise of online classes (such as massive open online courses or MOOCs) and hybrid online courses – are the direct result of an increased demand for access to education (Kena 2016). The imminent growth in enrollment to degree-granting institutions (Kena 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 impacts achievement gaps for female and URM undergraduates.

We address three questions by focusing on performance gaps between male and female students, and URM and majority students: 1) Does class size influence performance on exams? 2) Does class size influence performance on non-exam 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 received data from California State University, Chico; Cornell University; University of Minnesota, Twin Cities; and University of Puget Sound. We compared (1) pooled exam grades, (2) pooled assessments of student knowledge other than exams (hereafter non-exam grades; e.g., discussion sections, laboratories, online activities, written assignments, low-stakes quizzes, as well as active learning in-class activities), and (3) 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.
Figure 1. Four universities participated in the current study, representing diverse geographic locations across the US. Circle sizes are proportional to the number of classes sampled from each institution.
**Statistical Analyses**

**Linear mixed-effects model**

We used linear mixed-effects models to compare exam performance, performance on non-exam 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 a student outside of their class, as 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 students in biology classes at another university. For this reason, we use multilevel modeling to account for the non-independence of data in nested-data structures (Kreft et al. 1998, Paterson and Goldstein 1991).

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 based on this estimation using AIC differences ($\Delta i = AIC_{model} - \text{minAIC}$, where minAIC is the model with the smallest AIC value). Models with a $\Delta i > 10$ are considered poor predictors compared to the best model, and so we only present results with small $\Delta i$ values for brevity (Table S1). 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, URM status*class size).

In addition, we tested whether the following variables improved the fit of the model for the given set of data: (1) an interaction between student gender identity and URM status (SGender*URM status); (2) 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); (3) an interaction between student gender identity and instructor gender identity (SGender*IGender); (4) an interaction between student gender identity, URM status, and class size (SGender*URM status*class size); (5) 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 backwards eliminated from the final models ($P > 0.25$; see results). We used Bonferroni corrected post-hoc pairwise comparisons to clarify performance outcomes of students based on gender and URM status. We performed all statistical analyses using SPSS software version 24 (SPSS Inc., Chicago, IL, USA).

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model: Combined exam grades</th>
<th>AIC</th>
<th>Δi</th>
<th>Relative likelihoods</th>
<th>w_i</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>URM status + class size + SGender + class size*SGender</td>
<td>4885.468</td>
<td>0.000</td>
<td>1.000</td>
<td>0.935</td>
</tr>
<tr>
<td>2</td>
<td>URM status + class size + SGender + class size*SGender + age</td>
<td>4891.961</td>
<td>6.493</td>
<td>0.039</td>
<td>0.036</td>
</tr>
<tr>
<td>3</td>
<td>URM status + class size + SGender + class size<em>SGender + SGender</em>URM status + age</td>
<td>4892.347</td>
<td>6.879</td>
<td>0.032</td>
<td>0.030</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model: Non-exam grades</th>
<th>AIC</th>
<th>Δi</th>
<th>Relative</th>
<th>w_i</th>
</tr>
</thead>
</table>


<table>
<thead>
<tr>
<th>Rank</th>
<th>Model: Final course grade</th>
<th>AIC</th>
<th>Δi</th>
<th>Relative likelihoods</th>
<th>wi</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>URM status + SGender</td>
<td>4826.220</td>
<td>0.000</td>
<td>1.000</td>
<td>0.926</td>
</tr>
<tr>
<td>2</td>
<td>URM status + class size + SGender</td>
<td>4831.668</td>
<td>5.448</td>
<td>0.066</td>
<td>0.061</td>
</tr>
<tr>
<td>3</td>
<td>URM status + class size + Sgender + age</td>
<td>4836.220</td>
<td>10.000</td>
<td>0.007</td>
<td>0.006</td>
</tr>
<tr>
<td>4</td>
<td>Sgender</td>
<td>4837.260</td>
<td>11.040</td>
<td>0.004</td>
<td>0.004</td>
</tr>
<tr>
<td>5</td>
<td>URM status + class size + SGender + class size* Sgender + URM status</td>
<td>4837.736</td>
<td>11.516</td>
<td>0.003</td>
<td>0.003</td>
</tr>
</tbody>
</table>

Table 1. Best models for predicting performance metrics across four universities using AIC model selection. Compared to the first model, models with an Δi > 10 are considered poor predictors and so we do not report them here. The Akaike weights, wi, represent probabilities that a given model is the best model under repeated sampling.
Figure 2. The effects of class size on exam grade $z$-scores, non-exam grade $z$-scores, and final course grade $z$-scores for women (solid line) and men (dashed line). Colors represent different universities: University of Puget Sound (yellow), California State University, Chico (red), University of Minnesota, Twin Cities (purple), and Cornell University (blue).
Results

We used mixed model analyses to compare students’ combined exam grade, non-exam grade, and total course grade in the fall 2016 semester (Figure 2, 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 $B = 0.417, t(1377) = 6.01, P < 0.001, SE = 0.069$; Non-exam grade $B = 0.262, t(1533) = 3.83, P < 0.001, SE = 0.069$; Final course score $B = 0.407, t(1522) = 5.87, P < 0.001, SE = 0.069$). These results suggest that URM students’ exam scores on average was 0.42 standard deviation lower than non-URM students, and their non-exam scores were on average 0.26 standard deviation lower than non-URM students. Bonferroni corrected post-hoc pairwise comparisons, presented from the final models, show URM students underperforming on all performance metrics compared to non-URM students (Table 2; hereafter ‘underperform’ is used to describe raw gaps, and not those for which some measure of student academic ability/preparation is controlled). Second, we observed a significant interaction between gender and class size, such that as class size increased, women underperformed on exams ($S\text{Gender} \times \text{class size } B = -0.145, t(1599) = -2.89, P = 0.004, SE = 0.050$; Figure 2 inset) and in the course overall ($B = -0.108, t(1649) = -2.16, P = 0.031, SE = 0.050$) compared to men. We also found that women obtained higher non-exam grades ($B = 0.217, t(1731) = 4.60, P < 0.001, SE = 0.047$) compared to men, regardless of class size.

Next, we explored whether women are underperforming on exams because they are higher stakes in larger classes; i.e., they account for a larger proportion of the grade. To investigate this, we examined the correlation between class size and the percentage of students’
final course grades that are from their performance on exams. We did not find a strong
correlation (Pearson correlation = -0.386; \( P = 0.126 \)). This result runs counter to what one would
expect due to 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 grade (Koester et al. 2016). Finally, to test whether our results are the same within one
institution, we isolated twelve lower division classes from the University of Minnesota that
varied in class size. In these classes, all exams had identical multiple choice format. We found
the same main results across assessment types within one institution as we observed across all
institutions (Tables S4-S6). Thus, 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.

<table>
<thead>
<tr>
<th>Class size</th>
<th>50</th>
<th>150</th>
<th>250</th>
<th>50</th>
<th>150</th>
<th>250</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combined exam grade</td>
<td>-0.285</td>
<td>-0.295</td>
<td>-0.305</td>
<td>0.140</td>
<td>0.130</td>
<td>0.120</td>
</tr>
<tr>
<td>Non-exam grade</td>
<td>-0.165</td>
<td>-0.165</td>
<td>-0.166</td>
<td>0.086</td>
<td>0.086</td>
<td>0.085</td>
</tr>
<tr>
<td>Total course grade</td>
<td>-0.262</td>
<td>-0.264</td>
<td>-0.266</td>
<td>0.132</td>
<td>0.130</td>
<td>0.128</td>
</tr>
</tbody>
</table>

Table 2. Least-squares means comparison of relative performance of students who differ based
on their racial minority status (URM or non-URM) in different class sizes (50 students, 150
students, or 250 students). Measures are standardized, and reflect performance relative to the
mean of the class; positive scores are students who overperformed in standard deviations from
the mean, and negative scores represent those who underperformed relative to the mean. Our
data indicate that URM students underperform across all metrics, compared to non-URM
students, but unlike female students, their performance is not affected by class size, suggesting
factors other than class size negatively influence URM student performance. Standard errors are
shown in parentheses.
One possibility is that the positive effects we observe from 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., Ballen et al. 2017, Haak et al. 2014). Using data collected for nine of the seventeen courses (Table S7), we used a linear regression to examine the relationship between class size and 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 (Figure S1; Pearson correlation = -0.72; \( P = 0.028 \)), such that 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; \( P = 0.645 \)). Class size likely influences the frequency in which students interact with their instructor, 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.

Discussion

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

While 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 non-exam points—especially in lower division classes that serve as gateway courses to students’ major field 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 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 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], plicker cards [Howell et al 2017], etc.) and low-stakes—or no-stakes—formative assessments (e.g., one-minute papers, worksheets, and concept maps; Angelo and Cross 1993).

Because in our dataset female students excelled at non-exam assessments of the course material regardless of class size, an alternative strategy to promote women in STEM may be to make non-exam scores a larger component of the final course grade (Koester et al. 2016). Recent work shows traditional exams do not accurately capture student mastery of the cognitive skills required to do science and exacerbate existing gaps in performance (Moneta-Koehler et al. 2017, Stanger-Hall 2012). Further, women are adversely affected by test anxiety, which in itself is higher in women than in their male counterparts (Ballen et al 2017). Thus, 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 reinforce conceptual arguments against these classes (Achilles 2012, Baker et al. 2016, Glass 1982, Glass and Smith 1979, Ho and Kelman 2014, Schanzenbach 2014), 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 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 post-secondary
institutions is expected to increase 15 percent between 2014 and 2025 (Kena 2016). While 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. Thus, 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., pre-course measures such as SAT or cumulative GPA) for all
students across universities. Previous work finds that incoming preparation predicts performance
and retention across institutions (Ballen and Mason 2017, Ballen et al. 2017, Bonous-Hammarth
2000, 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. While our dataset is subject to some biases, these collaborative efforts among
universities allow for much larger datasets--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 gender gaps in performance that plague higher education.

Acknowledgements. We thank Daniel Baltz for help with data organization and interpretation; J.D. Walker and Lauren Sullivan for statistical support; 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: 1405E50826; CU: 1410005010; UPS: 1617-006; CSU: 461).
References


Clance PR. 1985. The impostor phenomenon: Overcoming the fear that haunts your success. Peachtree Pub Ltd.


Ventura J. 2000. I ain't got time to bleed: Reworking the body politic from the bottom up.
