Enhancing Diversity in Undergraduate Science: Self-Efficacy Drives Performance Gains with Active Learning

Cissy J. Ballen,[†‡*] **Carl Wieman,**[§‖] **Shima Salehi,**[§] **Jeremy B. Searle,**[†] **and Kelly R. Zamudio**[†]

[†]Department of Ecology & Evolutionary Biology, Cornell University, Ithaca, NY 14853; [‡]Department of Biology Teaching and Learning, University of Minnesota, Minneapolis, MN 55455; [§]Graduate School of Education and [‖]Department of Physics, Stanford University, Stanford, CA 94305

ABSTRACT

Efforts to retain underrepresented minority (URM) students in science, technology, engineering, and mathematics (STEM) have shown only limited success in higher education, due in part to a persistent achievement gap between students from historically underrepresented and well-represented backgrounds. To test the hypothesis that active learning disproportionately benefits URM students, we quantified the effects of traditional versus active learning on student academic performance, science self-efficacy, and sense of social belonging in a large (more than 250 students) introductory STEM course. A transition to active learning closed the gap in learning gains between non-URM and URM students and led to an increase in science self-efficacy for all students. Sense of social belonging also increased significantly with active learning, but only for non-URM students. Through structural equation modeling, we demonstrate that, for URM students, the increase in self-efficacy mediated the positive effect of active-learning pedagogy on two metrics of student performance. Our results add to a growing body of research that supports varied and inclusive teaching as one pathway to a diversified STEM workforce.

INTRODUCTION

As demographics in the United States become increasingly more diverse, we have committed to a number of efforts to improve the representation of historically underserved groups within science, technology, engineering, and mathematics (STEM) fields (e.g., Maton *et al.*, 2012; Wilson *et al.*, 2012; Hernandez *et al.*, 2013; Snyder and Wiles, 2015; Yeager *et al.*, 2016). Although critical to national interests (American Association for the Advancement of Science, 2011), progress in diversifying STEM is slow. One obstacle that underrepresented minority (URM) students face is overcoming a “chilly” classroom climate, characterized by little student participation and faculty-driven transmission of information in large introductory undergraduate classes (Hall and Sandler, 1982; Alexander *et al.*, 2009). This environment can significantly undermine students’ academic abilities and disproportionately affects historically URM students, who face unique challenges resulting from feelings of social isolation, low confidence, and stereotype threat (Steele and Aronson, 1995; Steele, 1997; Nguyen and Ryan, 2008). Interventions that specifically address these social–psychological issues have shown some success, but classroom environments must also impact student attitudes and consequent performance (Clewell, 1992; Cohen *et al.*, 2006; Walton and Cohen, 2011).

Recent studies also indicate that active-learning pedagogy (ALP)—characterized by in-class activities, prelecture preparation, and frequent low-risk assessment—increases

Erin L. Dolan, *Monitoring Editor*

Submitted December 9, 2016; Revised July 3, 2017; Accepted July 11, 2017

CBE Life Sci Educ December 1, 2017 16:ar56

DOI:10.1187/cbe.16-12-0344

Author contributions: C.J.B. conceived, designed, and performed the experiments. C.J.B. and S.S. analyzed the data. C.J.B., S.S., and K.R.Z. wrote the paper. C.J.B., C.W., S.S., J.B.S., and K.R.Z. edited and improved the manuscript.

*Address correspondence to: Cissy J. Ballen (balle027@umn.edu)

student learning and performance for all students (Freeman *et al.*, 2014) and often disproportionately benefits URM students and women compared with traditional lecture instruction (Lorenzo *et al.*, 2006; Beichner *et al.*, 2007; Freeman *et al.*, 2007; Haak *et al.*, 2011). However, underlying mechanisms leading to those benefits remain undemonstrated, although the conventional explanation is that students at risk need more structure in the educational environment (Haak *et al.*, 2011). Here, we propose another explanation: that ALP positively affects student well-being, which in turn enhances learning. To test this, we quantified student academic performance and two components of student well-being, self-reported confidence in the ability to do science, which we call science self-efficacy, and sense of social belonging (Walton and Cohen, 2011), in a large introductory STEM course that was modified from a traditional lecture format (n = 204; Fall 2014) to active instruction (n = 217; Fall 2015).

We addressed three specific questions: 1) Does ALP decrease the performance gap between non-URM and URM students? 2) Does ALP increase self-efficacy and perception of classroom social belonging? 3) Do these factors influence performance outcomes?

We chose science self-efficacy and classroom social belonging as two measures of well-being because of their demonstrated influence on student retention and performance in different educational contexts (Chemers *et al.*, 2011; Hurtado and Ruiz, 2012). We also expect that ALPs, such as increased interaction with instructors and among students, will directly impact well-being in the classroom and effectively encourage non-threatening interpretations of student interactions (Walton and Cohen, 2011). We also consider growth in elements of well-being as important stand-alone classroom outcomes.

MATERIALS AND METHODS

Quantifying Classroom Changes with Active Learning

Our study focused on an introductory evolutionary biology and biodiversity course (BioEE1780) at Cornell University that is required of all biology majors and attended primarily by students in their first year of college. In 2014, students came to class with no required preparation and listened to 50-minute traditional lectures with few interruptions or questions. In 2015, we implemented ALPs: 1) prelecture assignments (video podcasts and textbook readings); 2) low-risk prelecture quizzes; 3) assigned student groups working on structured problems in which students expressed their reasoning and worked together to solve problems during lecture; 4) personal response systems used for graded multiple-choice questions; and 5) redistribution of point allocation to reward group work and ongoing preparation rather than exam performance exclusively. Prelecture quizzes and in-class group work accounted for 18% of the final grade in the active semester. We expected students to participate in class and evaluated their engagement by rewarding iClicker points if groups participated, and taking away points if a group called on by the random number generator did not respond. In the traditional semester, exams accounted for 60% of the grade, compared with 42% in the active semester. Two examples of full-class activities developed by an instructor (C.J.B.) of Cornell's evolutionary biology course are now published as active-learning modules to accompany the *Life: The Science of Biology* textbook (Sadava *et al.*, 2017). These modules offer instructors engaging approaches to teaching challenging concepts in introductory biology, such as calculating the Hardy-Weinberg equilibrium or interpreting phylogenetic trees. Other examples of active-learning exercises included interpreting graphs and tables from the primary literature, predicting the most effective life history strategy given a set of environmental scenarios, and using backward elimination to identify a clade to which an unidentified organism belongs. BioEE1780 includes three 50-minute lecture sessions and one 50-minute discussion section each week. The discussion sections, meetings of smaller groups of students (15–20 individuals) led by graduate teaching assistants, remained the same throughout the study.

Instructor Experience

The instructors who participated in this collaboratively taught course each had at least 5 years of experience teaching BioEE1780 and had been coteaching this course every semester since 2009. However, none had previous formal experience teaching in an active format in a large lecture classroom. Over both semesters, nine instructors shared in teaching modules of the course, which included the following topics: phylogenetics, biodiversity, adaptation and speciation, population genetics, macroevolution, and human evolution. To rule out the possibility of instructor gender influence (Cotner *et al.*, 2011), both male and female instructors taught modules in each semester. All instructors received professional development training from the same active-learning postdoctoral associate (C.J.B.), which included guidance on developing and implementing activities that reached existing learning objectives for the course.

Student Demographics

In Fall 2014, the course was 60.7% female and 39.2% male; 35.9% Caucasian, 34.9% Asian American, and 21.4% URM (we defined URM students as those who are African American, Latino, Pacific Islander, and Native American, and non-URM students as those who are not underrepresented in STEM fields, including white students who are not of Hispanic origin and Asian-American students), with 8.1% of students declining to declare their ethnicity. In Fall 2015, the course was 55.7% female and 44.3% male; 38.2% Caucasian, 28.1% Asian American, and 25.4% URM, with 7.0% of students declining to declare their ethnicity. Active consent was collected from students each semester. We excluded four students over the two semesters who declined to participate in the study.

Data Collection

To compare student knowledge of course content across semesters, we used course grades and a pre–post knowledge assessment instrument (KAI). Because no research-validated concept inventory exists for use in our broad introductory evolutionary biology course, we designed the KAI to reflect the most important learning objectives listed in the syllabus and prelecture outlines (Supplemental Assessment S1). All nine faculty members who coinstruct the course contributed questions, edited, and approved the final KAI before its use. The KAI was distributed to students on the second day of class and then again on the last day of class and was not worth any grade points. We used Bloom's taxonomy (Bloom, 1956) to design questions for the KAI that reflected the level of learning we expected of students. This taxonomy identifies six levels of understanding: 1) knowledge,

2) comprehension, 3) application, 4) analysis 5) synthesis, and 6) evaluation. Level 1 relies on lower-order cognitive skills, such as memorizing concepts; higher levels require higher-order thinking, such as applying information in a new situation that is similar to the situation in which they learned it. We were interested in testing higher-order learning gains and thinking skills of URM students ($N = 58$) and non-URM students ($N = 196$) who completed the pre- *and* postcourse KAI. We asked two education experts in the Center of Teaching Excellence at Cornell University to assign a value of 1 to 6 to each KAI question. Ratings were performed separately, and we found substantial agreement between raters for both assessments (Cohen's kappa > 0.95). We computed a simple average of the ratings for each question (Supplemental Figure S1). Because the pre–post KAI is not worth points and is collected after each completion in class, there is no reason to think that students would ever have retained the assessment for their own later use or for the use of other students.

To examine the extent to which students felt confident comprehending, critically assessing, and communicating scientific concepts, and following Bandura's (Bandura, 1997) work on self-efficacy, we modified survey questions from an existing instrument (Robnett *et al.*, 2015) in which students rated confidence in their ability to complete course-relevant tasks. Responses were quantified on a five-point Likert scale (Supplemental Assessment S2): 1 = not confident; 2 = a little confident; 3 = somewhat confident; 4 = highly confident; and 5 = extremely confident.

We conducted principle component analyses on the six science self-efficacy survey items and three classroom-specific social belonging items. For science self-efficacy, we had adequate sampling to produce reliable results according to the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy for the whole data set (for presemester and postsemester results, KMO > 0.8). To test the presence of relationships between variables, we used Bartlett's test of sphericity, which we found to be significant (for presemester and postsemester results, $p < 0.001$). The precourse survey results generated a single component that explained 46% of the total variance; the postcourse surveys generated a single component that explained 56% of the total variance. We tested for internal consistency using Cronbach's alpha and found survey items to be correlated (Cronbach's alpha > 0.7). We then generated a single science self-efficacy response variable for each student by combining scores using an additive scale. We estimated the science self-efficacy differential among students by subtracting their precourse self-efficacy score from their postcourse self-efficacy score. We modeled the science self-efficacy gains with the combined precourse self-efficacy measure as a covariate to account for variation in incoming attitudes of students.

To examine sense of social belonging of students, we used four survey questions modified from Cornell University's Student Engagement and Inclusion Survey conducted by the University's Institutional Research and Planning group and in consultation with the University Diversity Council (for more information see http://irp.dpb.cornell.edu); these responses were also quantified on a five-point Likert scale (Supplemental Assessment S3). We asked students to what extent they agree or disagree with statements related to classroom and university social belonging, and the scale was coded as follows: 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = strongly agree. Students took the social belonging survey only once (at the end of the course), because it was designed to gauge social belonging over the entire semester. We combined three social belonging survey items, because they were specific to the classroom environment, and compared students' responses on the fourth item alone, because it gauged students' perceptions about the institution more broadly.

To test whether these data were suitable for factor reduction, we conducted a principal component analysis. For social belonging, the KMO measure of sampling adequacy for the whole data set was KMO = 0.619, and Bartlett's test of sphericity was $p < 0.001$. The three survey items generated a single component that explained 63% of the total variance. We tested for internal consistency using Cronbach's alpha, and found the survey items to be highly correlated (Cronbach's alpha > 0.8). In response to these results, we combined measures using an additive scale that represented a comprehensive classroom social belonging score for analyses.

Statistical Analysis

General Linear Analyses. We performed all statistical analyses using SPSS software version 24 (SPSS, Chicago, IL). We first used univariate general linear models to compare metrics of student achievement and well-being across the two semesters: learning gains (semester grade and gain in KAI), science self-efficacy, and social belonging in the classroom. For all analyses, we used Pearson correlations to examine whether baseline estimates (data collected before the course) were correlated with each other and with student outcomes. To fit the assumptions of the general linear model, we transformed students' grades by taking the linear log of [120 – student grade]. Owing to the presence of outliers in the residuals in our analysis of students' grades, we reran the analyses with the outliers excluded to make sure our findings were robust. The results were similar, and so the model presented here includes those outliers. For all Likert-scale analyses, we treated the dependent variables as continuous for ease of interpretation, given that nonparametric tests have yielded very similar results to the ones reported in this paper (Norman, 2010; Murray, 2013).

To analyze learning gains and well-being, we included semester (traditional or active), gender (female or male), URM status (URM or non-URM), and the interaction between semester and gender and semester and URM status as factors in all analyses (Supplemental Tables S1, S2, and S3). We also included students' incoming Scholastic Aptitude Test (SAT) math scores and precourse KAI scores and, as covariates in the analysis, the course grades and KAI scores, respectively. These two covariates were added to account for variation in students' incoming preparation for the course. An ANOVA showed that incoming math SAT scores (for non-URM students $F(1, 26) = 0.007$ $p = 0.933$; for URM students $F(1, 4524) = 1.064$ $p = 0.305$) and prelecture KAI scores (for non-URM students $F(1, 26.20) = 2.13$ $p = 0.145$; for URM students $F(1, 3.52) = 0.237$ $p = 0.628$) did not differ significantly between semesters, indicating that incoming student populations were comparable in their preparation. We also included presemester science self-efficacy score as a fixed effect in the analysis of science self-efficacy gains over the semester. We assessed model significance based on Akaike's information criterion (AIC). AIC allows us to estimate the best model for our data, based on an estimation using AIC differences (Akaike, 1974).

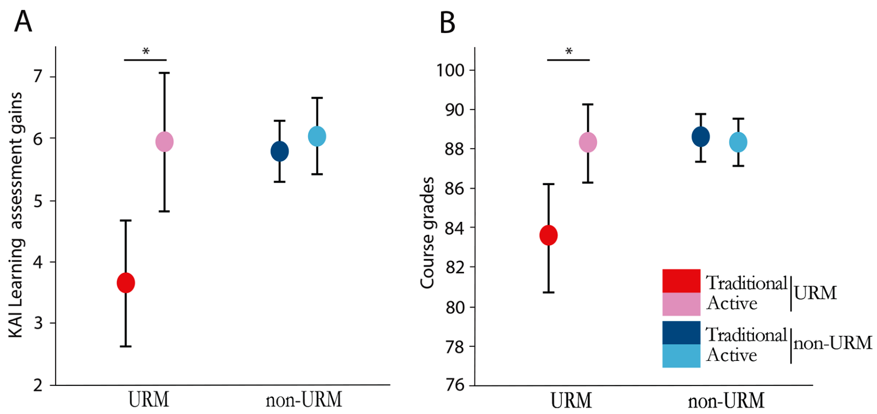

FIGURE 1. URM and non-URM student changes in academic performance for traditional and ALP courses. (A) Mean student learning gains (95% confidence interval) on the KAI, a 30-point assessment of course content. (B) Mean semester grades (95% confidence interval) controlling for incoming academic preparation. (Uncorrected means are 2015 active: URM = 86.35, SE = 0.97, *N* = 60; non-URM = 87.94, SE = 0.76, *N* = 157; 2014: URM = 80.02, SE = 1.86, *N* = 42; non-URM = 88.33, SE = 0.42, *N* = 162).

Mediation Analyses. Using separate full-mediation analyses, we tested the effects of pedagogy and student characteristics (gender and incoming preparation) on student performance, and whether performance gains were mediated by changes in scientific self-efficacy and sense of social belonging (see the Supplemental Material for detailed methodology).

RESULTS

Does Active Learning Decrease the Performance Gap between Non-URM and URM Students?

In 2014, non-URM students had significantly higher grades and KAI scores than URM students (Bonferroni post hoc pairwise comparison for both, $p < 0.0001$). This difference in performance disappeared in 2015 (course grades, $p = 0.938$; KAI gains, $p = 0.882$; Figure 1 and Supplemental Table S1).

Does Active Learning Change Students' Perception of the Class and of Their Abilities?

Reported self-efficacy increased from 2014 to 2015 equally for all demographic groups ($F(1, 1.73) = 6.55$, $p = 0.011$; Figure 2 and Supplemental Tables S2, S4, and S5). Classroom social belonging (Supplemental Tables S3 and S4) also increased significantly with ALP ($F(1, 2.47) = 4.20$, $p = 0.041$), but only for non-URM students (Supplemental Tables S3 and S6). However, there was no semester change in the degree to which students believed that Cornell demonstrates a commitment to diversity ($F(1, 0.18) = 0.173$, $p = 0.678$). This suggests that it is the classroom environment that changed and not the general perceptions of the student cohorts.

What Factors Influence Performance Outcomes?

Our previous analyses demonstrated that, although all students gained science self-efficacy in the active semester, non-URM students' academic performance metrics did not increase, while URM students' performance metrics significantly increased (Figure 1 and Supplemental Table S1). In light of these results, we concluded that further investigation was required and conducted mediation analyses after splitting the student sample according to minority status.

First, we compared the fit of partial- and full-mediation models with increase in self-efficacy being the mediating factor (Supplemental Figure S2). Semester, gender, and incoming academic preparation (incoming SAT math scores or precourse KAI scores) were covariates in the analyses of the course grades and KAI gains. With full mediation, the covariates predicted increased self-efficacy, which in turn predicted performance measures. The partial model included both this indirect mediating effect of covariates on performance plus the direct effect of covariates. For both grades and KAI, adding the direct effect of any of the covariates on performance did not improve the prediction compared with only having the indirect mediating effect of self-efficacy. For grades, the inclusion of the direct effect did not improve the fit (χ^2 (4) = 6.17, $p = 0.19$); for KAI, the full mediation with no direct effect fitted the data significantly better (χ^2 (4) = 13.7, $p = 0.008$). Therefore, full mediation is a better fit and more parsimonious model.

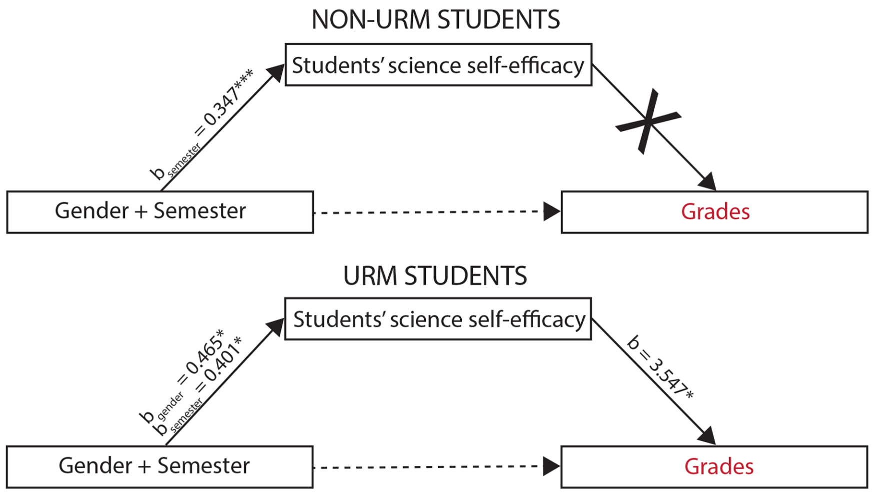

FIGURE 2. Analyses of non-URM and URM students show the mediation effect of self-efficacy on course grades (solid arrows) for URM students but no mediation for non-URM students. The partial-mediation model is illustrated by the dashed-line arrow. It tests the direct effects of pedagogy and student characteristics on performance and their indirect effect via scientific self-efficacy. In addition to the significant effects illustrated above, incoming academic preparation (e.g., SAT math score) also predicted all performance outcomes. *, $p \le 0.05$; ***, $p \le 0.001$.

For both grades and the KAI, the full-mediation models showed different results for non-URM and URM students. For non-URM, the covariates predicted self-efficacy, but self-efficacy was not correlated with performance (p_{grade} = 0.996; p_{KAI} = 0.685). For URM students, the covariates also predicted self-efficacy, but self-efficacy was correlated with performance, and in fact fully mediated the dependence of performance on the covariates (p_{grade} = 0.054; p_{KAI} = 0.010; Figure 2 and Supplemental Table S5). In other words, ALP increased self-efficacy, but this led to improved academic performance only for the URM students—an improvement that eliminated the learning gap.

Second, we carried out a similar mediation analysis for social belonging responses and found that social belonging was not significantly related to either performance metric across semesters (Supplemental Table S6). Although other studies have shown that an increased sense of belonging improves performance for URMs (Walton and Cohen, 2007, 2011), it is likely that this is not evident here, because the differences are small and more distal to class performance than is the self-efficacy measure.

DISCUSSION

Our data show that ALP improved knowledge of course material and that URM students benefited disproportionately. An active classroom using structured group activities also resulted in increased self-reported student confidence in scientific ability and overall increased classroom social belonging. Our analyses revealed that, for non-URM students, there was no mediation effect of science self-efficacy on performance. Conversely, for URM students, the increased science self-efficacy students experienced during the active-learning semester mediated the improved course performance (grades) and KAI gains. In other words, ALP increased students' science self-efficacy, and this led to improved academic performance for URM students. However, there was no such mediation effect for non-URM students. These results shed light on one mechanism that may underlie the positive effects of active-learning practices on URM students. Overall, our findings indicate that instructor efforts to incorporate active learning into their curricula can have positive results over the course of one semester.

This work has a few limitations that warrant consideration. First, we only compare cohorts of students across two semesters. While our work adds to compelling existing evidence that active learning benefits URM students (Beichner *et al.*, 2007; Freeman *et al.*, 2007; Haak *et al.*, 2011), replications of the current study are required to clarify the relationship between science self-efficacy, pedagogy, and performance. A longitudinal study design could address lasting impacts of ALP in introductory science courses. We may expect positive lasting impacts, particularly for URM students, who cite negative experiences in introductory science courses as the primary reason for declining interests in obtaining a science degree (Barr *et al.*, 2008). Second, we were unable to disaggregate URM student groups in the mediation analysis, because we would not have enough subjects to achieve adequate power to test for mediation. With a larger sample size, future work will be able to test the generality of these results and illuminate the impact of many affective measures on different URM groups after exposure to active learning.

Many studies support the notion that better pedagogy can lead to learning gains (Armbruster *et al.*, 2009; Haak *et al.*, 2011; Freeman *et al.*, 2014). A significant gap in the literature is the mechanism by which these gains occur, and why they benefit students in different demographic groups. Our results indicate that elements of classroom climate that promote collaborative problem solving, enhance group development, and engender confidence likely play an important role in learning. Instructors and researchers will profit from a deeper examination of other underlying mechanisms that impact achievement and well-being in underserved groups. For example, one characteristic feature of an active classroom is decreased reliance on a few high-stakes exams as primary contributors toward final course grade. Instead, the active classroom may also reward ongoing participation, in-class assignments, and group work. Future research should address the effects of exams and mixed assessment methods on students' well-being and course performance in traditional and active settings. Other benefits of active learning may result from the mediating effects of affective measures that we did not test here, such as engagement, motivation, and interest in course content.

Our findings underscore that students from different demographic groups may benefit in different ways from evidence-based teaching methods that emphasize interactive course design and collaboration. These teaching methods can reduce particular barriers that are faced by historically underrepresented students in STEM. The widespread adoption of these ALPs will be essential to our national efforts to improve diversity in STEM disciplines, while providing benefit to all students.

ACKNOWLEDGMENTS

We thank P. Lepage, L. Sanfilippo, A. Godert, E. Balko, and Cornell's Office of Undergraduate Biology and Center for Teaching Excellence for course support; F. Vermeylen for statistical advice; M. Clarkberg for help with student data; and H. W. Greene, M. L. Smith, L. Sanfilippo, A. Godert, G. Trujillo, and the Wieman lab for comments on the manuscript. This work was approved by IRB protocol number 1410005010 and funded by the College of Arts and Sciences, Cornell University.

REFERENCES

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*(6), 797–802.

American Association for the Advancement of Science. (2011). *Vision and change in undergraduate biology education: A call to action*. Washington, DC.

Armbruster, P., , Patel, M., Johnson, E., & Weiss, M.(2009). Active learning and student-centered pedagogy improve student attitudes and performance in introductory biology. *CBE—Life Sciences Education*, *8*(3), 203–213.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: W. H. Freeman and Company.

Barr, D. A., Gonzalez, M. E., & Wanat, S. F. (2008). The leaky pipeline: Factors associated with early decline in interest in premedical studies among underrepresented minority undergraduate students. *Academic Medicine*, *83*(5), 503–511.

Beichner, R. J., Saul, J. M., Abbott, D. S., Morse, J. J., Deardorff, D., Allain, R. J., … Risely, J. S. (2007). The student-centered activities for large enrollment undergraduate programs (SCALE-UP) project. *Research-Based Reform of University Physics*, *1*(1), 2–39.

Bloom, B. S. (1956). *Taxonomy of educational objectives. Vol. 1: Cognitive domain* (pp. 20–24). New York: McKay.

Chemers, M. M., Zurbriggen, E. L., Syed, M., Goza, B. K., & Bearman, S. (2011). The role of efficacy and identity in science career commitment among underrepresented minority students. *Journal of Social Issues*, *67*(3), 469–491.

Clewell, B. C. (1992). *Breaking the barriers: Helping female and minority students succeed in mathematics and science* (Jossey-Bass Education Series). San Francisco: Jossey-Bass.

Cohen, G. L., Garcia, J., Apfel, N., & Master, A. (2006). Reducing the racial achievement gap: A social-psychological intervention. *Science*, *313*(5791), 1307–1310.

Cotner, S., Ballen, C., Brooks, D. C., & 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.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences of the USA*, *111*(23), 8410–8415. doi: 10.1073/pnas.1319030111

Freeman, S., O'Connor, E., Parks, J. W., Cunningham, M., Hurley, D., Haak, D., … Wenderoth, M. P. (2007). Prescribed active learning increases performance in introductory biology. *CBE—Life Sciences Education*, *6*(2), 132–139. doi: 10.1187/cbe.06-09-0194

Haak, D. C., 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

Hall, R. M., & Sandler, B. R. (1982). *The classroom climate: A chilly one for women*? Washington, DC: Association of American Colleges and Universities.

Hernandez, P. R., Schultz, P., Estrada, M., Woodcock, A., & Chance, R. C. (2013). Sustaining optimal motivation: A longitudinal analysis of interventions to broaden participation of underrepresented students in STEM. *Journal of Educational Psychology*, *105*(1), 89.

Hurtado, S., & Ruiz, A. (2012). The climate for underrepresented groups and diversity on campus. *American Academy of Political and Social Science*, *634*(1), 190–206.

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

Maton, K. I., Pollard, S. A., McDougall Weise, T. V., & Hrabowski, F. A. (2012). Meyerhoff Scholars Program: A strengths-based, institution-wide approach to increasing diversity in science, technology, engineering, and mathematics. *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine*, *79*(5), 610–623.

Murray, J. (2013). Likert data: What to use, parametric or non-parametric? *International Journal of Business and Social Science*, *4*(11), 258–264.

Nguyen, H-H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology*, *93*(6), 1314.

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education: Theory and Practice*, *15*(5), 625–632.

Robnett, R. D., Chemers, M. M., & Zurbriggen, E. L. (2015). Longitudinal associations among undergraduates' research experience, self-efficacy, and identity. *Journal of Research in Science Teaching*, *52,* 847–867.

Sadava, D. E., Hillis, D. M., Heller, H. C., & Hacker, S. D. (2017). *Life: The science of biology*. London, UK: Macmillan Learning.

Snyder, J. J., & Wiles, J. R. (2015). Peer led team learning in introductory biology: Effects on peer leader critical thinking skills. *PLoS One*, *10*(1), e0115084.

Steele, C. M. (1997). A threat in the air. How stereotypes shape intellectual identity and performance. *American Psychologist*, *52*(6), 613–629.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*(5), 797–811.

Walton, G. M., & Cohen, G. L. (2007). A question of belonging: Race, social fit, and achievement. *Journal of Personality and Social Psychology*, *92*(1), 82.

Walton, G. M., & Cohen, G. L. (2011). A brief social-belonging intervention improves academic and health outcomes of minority students. *Science*, *331*(6023), 1447–1451. doi: 10.1126/science.1198364

Wilson, Z. S., Holmes, L., Sylvain, M. R., Batiste, L., Johnson, M., McGuire, S. Y., … Warner, I. M. (2012). Hierarchical mentoring: A transformative strategy for improving diversity and retention in undergraduate STEM disciplines. *Journal of Science Education and Technology*, *21*(1), 148–156.

Yeager, D. S., Walton, G. M., Brady, S. T., Akcinar, E. N., Paunesku, D., Keane, L., … Dweck, C. S. (2016). Teaching a lay theory before college narrows achievement gaps at scale. *Proceedings of the National Academy of Sciences of the USA*, *113*(24), E3341–E3348.