Context Matters: Social Psychological Factors That Underlie Academic Performance across Seven Institutions

S. Salehi,[1] S. A. Berk,[2] R. Brunelli,[3] S. Cotner,[4] C. Creech,[5] A. G. Drake,[6] S. Fagbodun,[7] C. Hall,[8] S. Hebert,[4] J. Hewlett,[9] A. C. James,[10] M. Shuster,[10] J. R. St. Juliana,[6] D. B. Stovall,[11] R. Whittington,[7] M. Zhong,[2] and C. J. Ballen[2]*

[1]Graduate School of Education, Stanford University, Stanford, CA 94305; [2]Department of Biological Sciences, Auburn University, Auburn, AL 36849; [3]Biological Sciences Department, California State University, Chico, Chico, CA 95929; [4]Department of Biology Teaching and Learning, University of Minnesota, Minneapolis, MN 55455; [5]Department of Biology, Mt. Hood Community College, Gresham, OR 97030; [6]College of Arts and Sciences, Cornell University, Ithaca, NY 14853; [7]Biology Department, Tuskegee University, Tuskegee, AL 36088; [8]Department of Biological Sciences, University of New Hampshire, Durham, NH 03824; [9]Department of Science and Technology, Finger Lakes Community College, Canandaigua, NY 14424; [10]Department of Biology, New Mexico State University, Las Cruces, NM 88003; [11]College of Arts and Sciences, Winthrop University, Rock Hill, SC 29733

ABSTRACT

To enhance equity and diversity in undergraduate biology, recent research in biology education focuses on best practices that reduce learning barriers for all students and improve academic performance. However, the majority of current research into student experiences in introductory biology takes place at large, predominantly White institutions. To foster contextual knowledge in biology education research, we harnessed data from a large research coordination network to examine the extent of academic performance gaps based on demographic status across institutional contexts and how two psychological factors, test anxiety and ethnicity stigma consciousness, may mediate performance in introductory biology. We used data from seven institutions across three institution types: 2-year community colleges, 4-year inclusive institutions (based on admissions selectivity; hereafter, inclusive), and 4-year selective institutions (hereafter, selective). In our sample, we did not observe binary gender gaps across institutional contexts, but found that performance gaps based on underrepresented minority status were evident at inclusive and selective 4-year institutions, but not at community colleges. Differences in social psychological factors and their impacts on academic performance varied substantially across institutional contexts. Our findings demonstrate that institutional context can play an important role in the mechanisms underlying performance gaps.

INTRODUCTION

Broadening participation in science, technology, engineering, and mathematics (STEM) fields requires acknowledging the many different student experiences in higher education. Students who identify with groups historically excluded from STEM fields face barriers as they navigate these pathways into the STEM workforce (Seymour and Hewitt, 1997; Seymour and Hunter, 2019). Social challenges that disproportionately affect such students include social exclusion (Hurtado and Ruiz, 2012), a negative campus climate (Koo, 2021), and discrimination from peers and authority figures (Hurtado and Ruiz Alvarado, 2015; Park *et al.*, 2020). Meanwhile, national efforts call for more STEM graduates to address the shortage of STEM

Rebecca Price, *Monitoring Editor*
Submitted Jan 20, 2021; Revised Sep 15, 2021; Accepted Sep 29, 2021
CBE Life Sci Educ December 1, 2021 20:ar68
DOI:10.1187/cbe.21-01-0012
*Address correspondence to: C. J. Ballen (mjb0100@auburn.edu).

workers in the labor market and to expand underrepresented minority (URM[1]) participation (American Association for the Advancement of Science, 2011; National Academy of Sciences, National Academy of Engineering, and Institute of Medicine, 2011; President's Council of Advisors on Science and Technology, 2012). Increased participation of students in STEM education will directly impact national science literacy, provide high-quality education to the STEM workforce, and contribute to critical scientific advances. One significant barrier to student participation and persistence in STEM is introductory STEM courses that students are required to complete before moving on to more advanced course work (Crisp *et al.*, 2009; Mervis, 2011). In such classes, instructors deliver foundational content knowledge—often through undisrupted lecture (Stains *et al.*, 2018)—to relatively large lecture halls of students. Failure and attrition rates are high (Seymour and Hunter, 2019), and students perform poorly in these courses relative to non-STEM courses (Mervis, 2011; Koester *et al.*, 2016). URM students face additional barriers in these introductory courses, in large part due to gaps in incoming preparation (Harris *et al.*, 2020; Salehi *et al.*, 2020) and limited opportunities in some institutions to interact with instructors (Hurtado *et al.*, 2011). These barriers lead to underperformance and lower probability of persisting in STEM fields for URM students compared with their overrepresented majority (non-URM) peers (National Science Foundation, 2019; Salehi *et al.*, 2020).

To address student performance barriers in introductory biology courses, recent research in biology education focuses on best practices for evidence-based teaching strategies that improve student academic performance (Offerdahl *et al.*, 2011; Freeman *et al.*, 2014; Theobald *et al.*, 2020). Though important to our understanding of student learning, the majority of current research into student experiences in introductory biology takes place at large, predominantly White institutions, such as public, 4-year PhD- and master's degree–granting research-intensive universities (hereafter, *selective* institutions; Schinske *et al.*, 2017; Thompson *et al.*, 2020). This overrepresents the experiences of these students in the contemporary literature. Less research has focused on other types of institutions, such as community colleges, minority-serving institutions, or more inclusive 4-year institutions (hereafter, *inclusive* institutions), even though they are essential educational institutions contributing to our future STEM workforce (Thompson *et al.*, 2020). For example, while community colleges serve almost half of all undergraduate students in the United States, and an even larger percentage of URM and first-generation college students (American Association of Community Colleges [AACC], 2021), relatively little biology education research (BER) focuses on community colleges (Schinske *et al.*, 2017). Further, the authors who conduct research on community colleges are often not affiliated with community colleges and less familiar with the specific institutional context (Schinske *et al.*, 2017). The reason for this discrepancy is likely due to heavy teaching and service expectations for community college faculty, but still represents a gap in our understanding of the student experience.

By developing and implementing innovations and reforms informed primarily by research conducted at selective institutions, we overlook important contextual factors that likely differ between selective institutions and community colleges and inclusive institutions; factors such as instructor teaching loads and expectations, students' levels of academic preparation, their educational experience based on salient social identities such as gender or race/ethnicity, and students' socioeconomic status. Pursuing research questions across multiple institutional contexts enables research to address questions that are unattainable within a single institution. Multiple factors specific to an individual institution are likely to impact research outcomes. These could include (but are not limited to) teaching culture, student body, or class size. To demonstrate the importance of institutional context, we examined differences in demographic performance gaps, as well as two psychological factors, specifically test anxiety and ethnicity stigma consciousness (hereafter ESC; Sarason, 1961; Pintrich *et al.*,1993; Steele and Aronson, 1995; Picho and Brown, 2011). We tested their impacts on student performance across seven institutions from three different institution types: community colleges, 4-year inclusive institutions, and 4-year selective institutions. We selected these two social psychological factors because they are potentially impacted by institutional or classroom features (Matthews *et al.*, 1999; Ergene, 2003; Massey and Fischer, 2005; Osborne and Walker, 2006; Kellow and Jones, 2008; Ballen *et al.*, 2017b) and can have consequential impacts on learning and academic performance (Kellow and Jones, 2008; Appel and Kronberger, 2012; Salehi *et al.*, 2019a).

We selected test anxiety to measure across contexts because high levels of test anxiety reduce student performance outcomes, and we predicted this may vary across multiple contexts based on factors such as class size, course competitiveness, or other traits that may relate to institution type (von der Embse *et al.*, 2018). Test anxiety is characterized by feelings of tension, worried thoughts, and negative physiological reactions in an academic evaluative setting (American Psychiatric Association, 2013). Work across STEM fields shows that test anxiety negatively impacts assessment performance outcomes (Chapell *et al.*, 2005; DordiNejad *et al.*, 2011; Ali and Mohsin, 2013; Shapiro, 2014; Harris *et al.*, 2019). Results from previous research showed that both in introductory biology (Ballen *et al.*, 2017b; Cotner *et al.* 2020) and across STEM disciplines in higher education (Salehi *et al.*, 2019b), for women only, test anxiety negatively influenced exam performance. Work conducted at a separate institution could not replicate these results and concluded that women underperform on exams for reasons other than gender-based differences in test anxiety (Harris *et al.*, 2019). It is possible that students with higher content-related anxiety perceive test difficulty differently, which may in turn impact test anxiety (Hong and Karstensson, 2002). The impacts of test anxiety on performance of students with underrepresented identities in biology has not been studied as extensively, to our knowledge. One study conducted mediation analysis to examine the relationship between a number of social psychological factors, including test anxiety, as well as incoming preparation, minority and first-generation status, and academic performance outcomes

[1] We define "URM" as a category that comprises three racial or ethnic minority groups whose representation in science education or employment is smaller than their representation in the U.S. population according to the U.S. National Science Foundation definition (found here https://bit.ly/2BZx1ZO; Black or African Americans, Hispanics or Latinos, and American Indians or Alaska Natives). However, the definition of URM used by diversity programs in the United States varies (Page *et al.*, 2013), and we recognize that aggregating data into a URM category hides significant interracial inequalities (Bensimon, 2016).

within a college of biological sciences (Salehi *et al.*, 2020). Results showed that URM and first-generation status was highly related to students' incoming preparation. Lower measures of incoming preparation were associated with higher anxiety scores, and higher anxiety was associated with lower exam scores.

We also selected ESC to measure across contexts. ESC is a measure of stereotype threat, which we measured as the extent to which one feels self-conscious of the stigma associated with one's race/ethnicity (Picho and Brown, 2011). Stereotype threat can invoke a disruptive state, undercutting performance and aspirations within a domain (Spencer *et al.*, 2016). Ever since Steele and Aronson's (1995) work on this phenomenon, research has documented its presence among negatively stereotyped individuals in hundreds of studies. Stereotype threat has been shown to reduce an individual's ability to learn (Rydell *et al.*, 2010; Taylor and Walton, 2011; Boucher *et al.*, 2012). It can be reduced or invoked through relatively small changes in instructor behavior or the classroom environment, such as stating that tests are difficult and performance is linked to overall ability (Steele and Aronson, 1995). Little work on stereotype threat has been conducted in the context of postsecondary biology. Taasoobshirazi *et al.* (2019) administered study questions with different stereotype threat conditions in the instructions and examined biology performance based on gender, as well as self-efficacy, motivation, and domain identification. While they did not identify differences in gender outcomes by treatment, they observed women reported a greater domain identification with biology. To our knowledge, no previous work has focused on stereotype threat based on race/ethnicity in the context of postsecondary biology. Additionally, quantifying this social psychological factor as it varies across classrooms or institutions has not been explored, despite accumulating evidence that institutional context shapes student experiences (Chang *et al.*, 2008; Hurtado *et al.*, 2010; Wang, 2013; Winkle-Wagner and McCoy, 2018). For example, we may expect ESC to be higher in predominantly White spaces such as the selective institutions in our sample.

Using data generated collaboratively across multiple institutions, we tested the following research questions:

1. To what extent do we observe gaps in academic performance based on demographic status such as binary gender[2] and URM status in introductory biology across different institutional contexts?
2. To what extent do we observe differences in social psychological factors, that is, test anxiety and ESC, based on demographic status in introductory biology across different institutional contexts?
3. How do these social psychological factors mediate potential differences in performance outcomes based on demographic status across different institution contexts?

Traditional views of the performance gap attribute student success primarily to personal attributes, such as deficiencies in ability or motivation. Focusing on performance gaps in this way deflects attention away from the key responsibility educational institutions play in student success and applies underlying deficit perspectives to marginalized students in STEM (Smit, 2012; Zhao, 2016). Here we stress that the *opportunity* gap in primary and secondary education—for example, access to high-quality curricula, instruction, and technology—contributes to gaps in academic preparation before students even enter the undergraduate biology classroom. These gaps in academic preparation are then in turn reflected in student grades in introductory courses that fail to provide students with equal opportunities to perform (Salehi *et al.*, 2019, 2020). Understanding the extent of performance disparities in courses and their prevalence across institutional contexts is an important first step to fulfill the responsibility of higher education to educate all students. Our focus on academic disparities aims to change structural barriers within our classrooms and institutions such that student performance patterns (e.g., exam grades, total grades) are not predicted primarily by characteristics such as race, class, ethnicity, gender, proficiency in a dominant language, or other marginalized student identities.

METHODS

Participants

We solicited participation through an existing professional network from instructors who teach biology at a range of institutions (Research Coordination Network, National Science Foundation RCN–UBE Incubator: Equity and Diversity in Undergraduate STEM; grant no. 1729935 awarded to S.C. and C.J.B.; RCN–UBE grant no. DBI-1919462 awarded to S.C., C.J.B., S.F., Harshman, and C.H.). Equity and Diversity in Undergraduate STEM is a network of educators and discipline-based education researchers who work together on research and teaching projects in the context of biology curricula and is supported by the National Science Foundation. The objectives of the network are to: "(1) reveal differences, if they exist, in the cultural climate for women and minorities in STEM disciplines (initially focusing on biology) as a function of geography, institution type, and cultural profile of the participating departments; (2) increase the number of faculty in the United States that are familiar with barriers to inclusion in STEM, and can apply evidence-based techniques for countering known barriers; (3) develop a community of faculty that can serve as leaders-at their home institutions and nationally-in inclusive teaching and assessment; and (4) identify cultural factors associated with a shift toward evidence-based teaching, especially pertaining to inclusive teaching" (Thompson *et al.*, 2020).

From the RCN network, we collected data from 33 biology courses at seven institutions between 2016 and 2018. Participating institutions were a convenience sample chosen from a range of institutional types (public and private, large and small) and settings (college towns to large metropolitan areas). We classified these institutions based on undergraduate profile according to the Carnegie Classification of Institutions of Higher Education. This allowed us to group institutions by admissions selectivity criteria and by whether they were 2-year or 4-year institutions. The 2-year institutions in our sample were community colleges, and so use we this term throughout this research. Within 4-year institutions, we classified institutions as "selective" if they fell into the Carnegie classification of "more selective," indicating that the 25th percentile of first-time first-year

[2] A limitation of our research includes data that are not inclusive to transgender, nonbinary and/or gender-nonconforming people. This was due to low sample sizes in our sample, which can lead to student privacy concerns. We hereafter use the term "gender" to describe men and women, while acknowledging the limitations of these two categories and the need for future work to be more inclusive of the continuum of gender.

TABLE 1. Summary of participating institutions: each institution's average entrance exam scores (when applicable); whether each is a minority-serving institution (year of designation); the approximate number of undergraduates enrolled; how evidence-based teaching is promoted at the institution; pedagogy of participating classes; typical class size for lower-division (LD) classes at the institution; typical class size for upper-division (UD) classes at the institution; and typical teaching load of instructors who participated in the study[a]

Institution	Avg. SAT or ACT	MSI? (year)	Approx. no. of undergraduates	How is evidence-based teaching (e.g., active learning) promoted at your institution, if at all?	Pedagogy[b]	LD class size (i.e., first and second year)	UD class size (i.e., third and fourth year)	Teaching load[c]
Inclusive 4-year.1	SAT 950–1150; ACT composite score of 18–23	Yes (2009)	12,000	There are various peer-led initiatives funded internally and by grants and there is strong support in the teaching academy.	Active learning	150	75	3 to 5 classes per year
Inclusive 4-year.2	ACT composite score of 20–22	Yes (1881)	2000	One of the president's main focus areas is creating a student-centered campus environment. The university brings in speakers during all-university conference (twice a year) to discuss teaching strategies. The last few speakers spoke about active learning and how to engage students. We have a seminar once a semester called Strategies for Effective Engagement (SEE). The speakers have been many high school teachers with 20+ year experience on engaging students. However, the teaching method is up to the faculty member.	Interactive lecture	60	40	More than 5 classes per year
Inclusive 4-year.3	SAT 1000–1190; ACT composite score of 18–24	Yes (2016)	17,000	It is promoted in many different majors and student organizations/activities university-wide. Examples include: building competitions in engineering, the great debate, the hands-on lab for future teachers, sustainability conference, first-year experience courses, ecology labs at Big Chico Creek Ecological Reserve, nursing clinical labs, Community Action Volunteers in Education (CAVE), student-run university newspaper, and many more.	Active learning	120	40	More than 5 classes per year
Selective 4-year.1	SAT EBRW 680–750; SAT Math 720–790; ACT Composite score of 32–35	No	15,000	There is an active-learning initiative, started in 2013, that is funded by donors and gives out grants to departments that apply each semester for funding. About 25% of the departments at [Institution] now have an association with the active-learning initiative.	Active learning	300	40	1 to 3 classes per year
Selective 4-year.2	SAT 1080–1260; ACT composite score of 22–28	No	12,000	Few opportunities or incentives to learn about and use evidence-based teaching exist.	Traditional lecture	200	50	3 to 5 classes per year

(*Continues*)

TABLE 1. Continued

Institution	Avg. SAT or ACT	MSI? (year)	Approx. no. of undergraduates	How is evidence-based teaching (e.g., active learning) promoted at your institution, if at all?	Pedagogy[b]	LD class size (i.e., first and second year)	UD class size (i.e., third and fourth year)	Teaching load[c]
Selective 4-year.3	ACT composite score of 28	Yes (2016)	31,000	There are several institution-level initiatives (e.g., through the Center for Educational Innovation) as well as various collegiate and department-level programs to support professional development around making courses more "active." While there are pockets of resistance, these are not at the administrative level.	Interactive lecture	115	20	3 to 5 classes per year
Community. college.1	N/A SAT/ACT scores are not required or collected.	No	60,000	Few opportunities or incentives to learn about and use evidence-based teaching exist.	Traditional lecture	50	15	3 to 5 classes per year
Community. college.2	N/A SAT/ACT scores are not required or collected.	No	5,600	The Center for Teaching and Learning provides regular professional development opportunities.	Active learning	25	10	More than 5 classes per year

[a] Original reports edited for clarity and succinctness.
[b] Select: 1) traditional lecture; 2) interactive lecture: lectures with some opportunities for student interaction, such as instructors posing questions; or 3) active learning: lectures with iClickers, group work, formative assessments, and pre-readings.
[c] Select 0–1 class per year, >1 to 3 classes per year, >3 to 5 classes per year, or >5 classes per year.

students received greater than 21 on their American College Testing (ACT) entrance exam score. We classified institutions as "inclusive" if they fell into the Carnegie category of "inclusive" or "selective," with 25th percentile ACT scores less than 21. Additionally, we worked with representatives at each institution to generate Table 1, which summarizes some contextual characteristics of the participating institutions.

We worked with individual instructors who teach biology classes across these institutions to collect data. Our sample represented 12 courses at two community colleges (n = 454 students, average class size = 38 ± 6.68), six courses at three inclusive institutions that were also minority-serving institutions (n = 1045 students, average class size = 199 ± 49.85), and 15 courses at three selective institutions (n = 3594 students, average class size = 239 ± 40.26). During the 2-year study period, we obtained grade data disaggregated into exam scores, non-exam scores, and final grades and institutional data on gender and race/ethnicity from all courses in the study. Also, during the last few weeks of the semester for a subset of students, we administered a survey to measure test anxiety and ethnicity stigma consciousness (Table 2). Note we removed one institution, representing one class, from our analyses of demographic gaps due to the fact that 100% of student were from underrepresented racial minority groups. All aspects of research were reviewed and approved by each institution's institutional review board.

Survey Measures

To measure test anxiety, we used a four-item test anxiety construct from the Motivated Strategies for Learning Questionnaire (MSLQ; Pintrich *et al.*, 1993). The MSLQ test anxiety construct asked students about their experiences during testing, such as whether they felt distracted by their anxiety during exams. To measure ESC, we used the five-item ESC construct from the Social Identities and Attitudes Scale (SIAS; Picho and Brown, 2011). ESC measured the extent to which one is conscious about one's ethnic identity and includes items such as "My ethnicity influences how teachers interact with me." For both constructs, students rated their agreement with each item on a seven-point Likert scale ranging from "strongly disagree" to "strongly agree." We have included the full list of survey items in Supplemental Table S1.

Analysis

We performed all statistical analyses in R (R Core Team, 2019). We reported exam scores as the average of all exam scores earned by each student, and we normalized this average for each course to control for variation in grading and exam structure across courses. Therefore, each student's exam score can be interpreted as how many standard deviations that score is from the mean of a given course.

Validating Constructs: Confirmatory Factor Analysis. We used confirmatory factor analysis (CFA) in the R package lavaan to verify our survey item structure for test anxiety and ethnicity stigma consciousness. We used CFA because these survey items were previously developed and established as theoretical models and validated in previous research (Picho and Brown, 2011; Pintrich *et al.*, 1993; Rosseel, 2012). Thus, we were testing to confirm whether our data set supports the preexisting survey with established structure (Knekta *et al.*, 2019). For advantages

TABLE 2. Number of student responses to survey constructs, exam scores, and percentage of URM students for each institution type[a]

Institution type	Test anxiety construct (I/C)	Ethnicity stigma consciousness construct (I/C)	Exam scores (I/C)	% URM students ± SD (range)
2-year (community college)	342 (2, 12)	344 (2, 12)	454 (2, 12)	26% ± 13% (11–60%)
Inclusive 4-year	747 (2, 5)	512 (2, 4)	995 (2, 5)	50% ± 14% (36–65%)
Selective 4-year	845 (3, 6)	1937 (3,12)	3594 (3, 15)	10% ± 5% (2–25%)

[a] Numbers in parentheses represent number of institutions (I) and courses (C) for each measure. Percentage of URM students represents the average percentage of URM students in each class for each institution.

and more details of conducting CFA via lavaan in biology education, see Ballen and Salehi (2021). To evaluate how well the CFA model captured the variation in the survey data, we used structural equation modeling (SEM) fit indices: confirmatory fit index, root mean square error of approximation, SRMR. All the fit indices fell within the acceptable range, suggesting the survey items loaded on the defined factors and the CFA model properly captures the variation in the data.

Quantifying Performance Gaps: Mixed-Effect Analysis of Performance Gaps. To test the extent to which gaps in academic performance are related to demographic status such as gender and URM status (together referred to as "demographic status") in introductory biology, we ran mixed-effects multivariable regression analysis using the nlme package in R. We controlled for the random effect of courses nested within each institution.

Mediation Analysis. We also tested: 1) whether performance outcomes based on gender and URM status were mediated by test anxiety or ESC; and 2) whether mediation varied across institution type (community college, inclusive institutions, or selective institutions). To this end, we used a SEM approach using lavaan package in R (Rosseel, 2012) on a subset of students for whom we had responses to all survey items and their exam scores (n = 337 students in 12 courses at two community colleges, n = 581 students in five courses at two inclusive institutions, and n = 756 students in five courses at three selective institutions). In lavaan, we used group analysis, with institution type as the grouping factor. Therefore, the proposed mediation model would be fit to the data of each institution type separately, and whether the model was a good fit for each institution type would be tested. We tested for mediation by using the Akaike information criterion (AIC) to compare two structural models: one including only indirect effects of demographic status mediated by social psychological factors and one that also included the direct effect of demographic status on exam scores. Within the SEM analysis, we included the latent variable structure supported by our CFA. For ease of interpretation, we normalized these construct responses for the entire sample across institution types, such that reported values reflect how many standard deviations a value is from the sample's mean score.

RESULTS

Validating Constructs: CFA

We removed one of the survey items from the SIAS ("My ethnicity impacts how I interact with people of other ethnicities"), which had a poor loading in our sample (standardized factor loading <0.70). We evaluated model fit using several fit indices, and while RMSEA, SRMR, and CFI were consistent with cutoff values (Hu and Bentler, 1999; Knekta *et al.*, 2019), our chi-square test of model fit was highly significant (CFI = 0.99, RMSEA = 0.03, SRMR = 0.016, $\chi^2 = 50.91$, $p < 0.001$). However, this is likely an effect of our large sample size, because chi-square tests are highly sensitive to sample size. In large samples, even a very small difference is highly likely to be significant (Hu and Bentler, 1999). Therefore, with a large sample size, if the model estimated covariance matrix is slightly different from the observed covariance matrix, the chi-square statistics might be still significant. Given that all three other indices fell within the acceptable range, and our constructs had high reliability (α = 0.78; test anxiety(α) = 0.89; ESC(α) = 0.89) we accepted our CFA in lavaan as a good fit.

Quantifying Performance Gaps: Mixed-Effect Analysis of Performance Gaps

For gender performance analysis, we found that the nested structure of the random effect explained 4% of the variance in the data. This suggests that the random effects due to courses nested within institutions were small. This is partly due to the fact that the course grade is normalized within each course. We found no significant effect of gender on average exam scores (mixed-effects model: $\beta_{CC} = 0.09 \pm 0.11$, $p = 0.96$; $\beta_{inclusive} = 0.07 \pm 0.08$, $p = 0.93$; $\beta_{selective} = 0.08 \pm 0.03$, $p = 0.13$). While there was also variation in the extent of the gap between men and women, a greater percentage of courses showed no difference in exam scores by gender. For URM performance analysis, we similarly found that the nested structure of the random effects explained 4% of the variance in the data, which suggests that random effect is relatively small. We found that URM students underperformed relative to non-URM students on exams at both selective and inclusive 4-year institutions, but not at community colleges (mixed-effects model: $\beta_{CC} = -0.23 \pm 0.12$, $p = 0.30$; $\beta_{inclusive} = -0.27 \pm 0.07$, $p = 0.003$; and $\beta_{Selective} = -0.44 \pm 0.05$, $p < 0.001$). As we only found significant URM performance gaps in our sample and no gender gaps, we proceeded to only examine how social psychological factors mediate URM performance gaps.

Mediation Analysis

Our structural equation model revealed variation in mediation effects across institution types (Figure 1). We found support for partial mediation effects, as the partial mediation model was a better fit for the data ($p = 0.002$) and as the ΔAIC between the partial mediation and full mediation model was 9, which is larger than our cutoff ΔAIC of 2. In this model, test anxiety was universally negatively associated with exam scores ($\beta_{CC} = -0.27 \pm 0.02$, $p < 0.001$; $\beta_{inclusive} = -0.13 \pm 0.03$, $p = 0.007$; $\beta_{selective} = -0.33 \pm 0.02$, $p < 0.001$). However, only at inclusive institutions, URM students had higher test anxiety. For community colleges or selective institutions, there was no difference in test anxiety across URM status (β_{CC} = 0.16 ± 0.24, p = 0.21; $\beta_{inclusive}$ = 0.30 ± 0.16, p = 0.001; $\beta_{selective}$ = 0.18 ± 0.23, p = 0.21). These relations imply that test anxiety mediated the effect of URM status on exam scores *only* at the inclusive institutions, but not at community colleges or selective institutions. The size of the mediation path can be calculated by multiplying the coefficients included in each path (β_{CC} = –0.04 ± 0.04, p = 0.22; $\beta_{inclusive}$ = –0.04 ± 0.02, p = 0.04; $\beta_{selective}$ = –0.06 ± 0.05, p = 0.22). In other words, URM students at inclusive institutions had higher test anxiety than their non-URM peers, and this higher test anxiety led to lower exam performance. At community colleges and selective institutions, on the other hand, there was no difference across URM status in test anxiety.

ESC was not a significant mediator for the effect of URM status on exam scores for all the three institution types (β_{CC} = 0.01 ± 0.01, p = 0.41; $\beta_{inclusive}$ = –0.01 ± 0.01, p = 0.26; $\beta_{selective}$ = –0.008 ± 0.02, p = 0.72). However, the reason for the lack of mediation effect varied by institution type. We did not observe an effect of URM status on ESC at community colleges (β = 0.12 ± 0.15, p = 0.38), but ESC positively affected exam scores (β = 0.1 ± 0.04, p = 0.02). At inclusive institutions, there was no significant correlation of URM status with ESC (β = 0.15 ± 0.12, p = 0.12), and no correlation of ESC with exam scores (β = –0.07 ± 0.04, p = 0.11). However, we did observe a correlation of URM status with ESC at selective institutions (β = 0.57 ± 0.13, p < 0.001) similar to the previous work (Salehi *et al.*, 2020), though there was still no effect of ESC on exam scores (β = –0.01 ± 0.04, p = 0.72). The observation that, for selective universities, none of social psychological factors are mediators for URM performance gaps aligns with previous findings (Salehi *et al.*, 2020) that the incoming preparation was the major mediator for demographic performance gaps in an R1 institution, and after controlling for that effect, the social psychological factors did not have a mediating effect for demographic performance gaps.

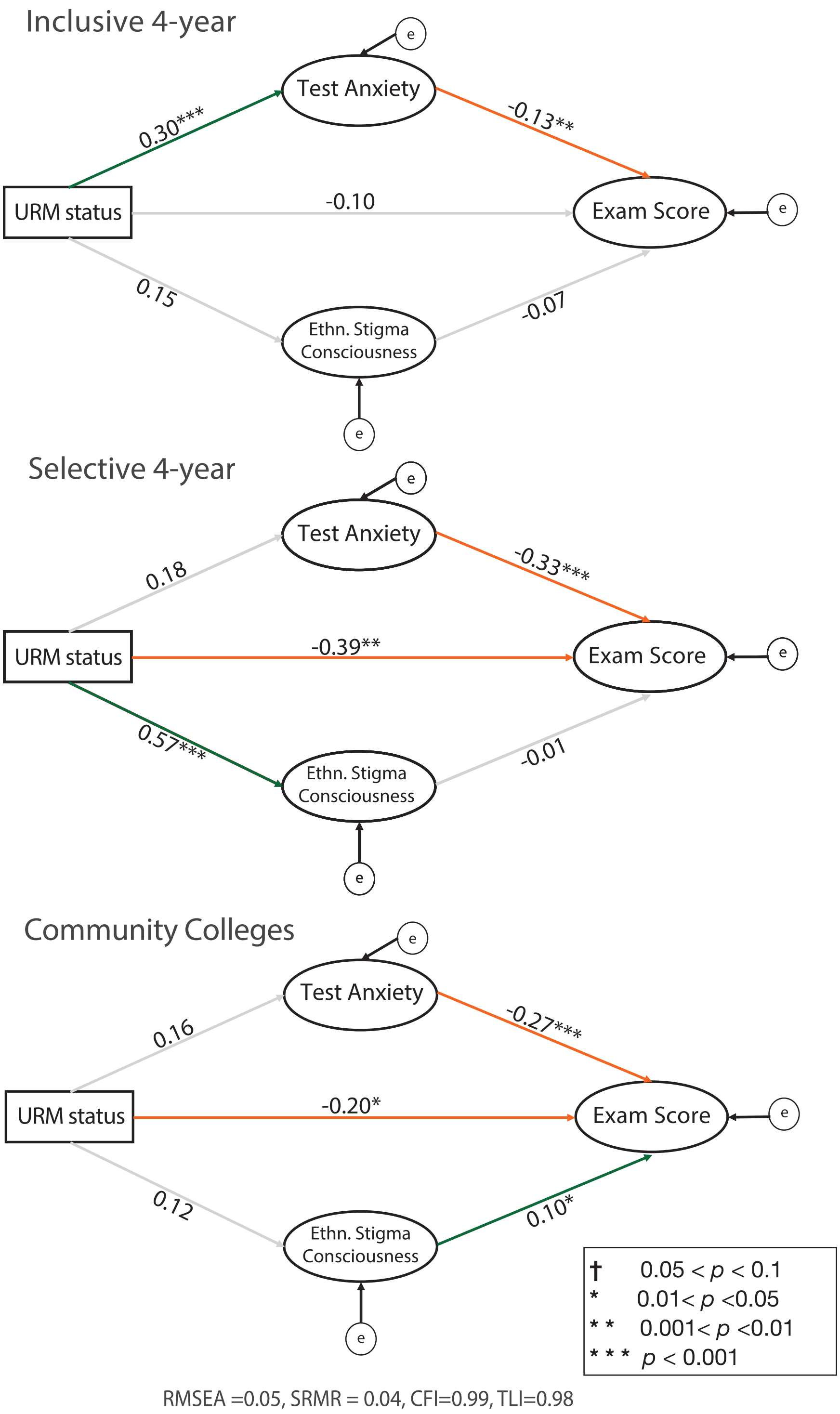

FIGURE 1. Structural equation model for mediation effects of test anxiety and ethnicity stigma consciousness across institution types. All continuous variables have been rescaled to have a mean of 0 and an SD of 1 and can be interpreted as positive or negative units of SD reflecting the relationship between two variables. An "e" in a circle indicates an error term in our estimations of a model variable. Orange lines indicate negative correlations; green lines indicate positive correlations. For example, the figure implies that, for inclusive institutions, URM students on average have 0.3 SD higher test anxiety compared with non-URM students, and a 1 SD increase in text anxiety leads to a 0.13 SD decrease in exam score. The size of each mediation path is calculated by multiplying the size of each coefficient included in that path. Hence, the size of mediation effect for test anxiety in inclusive institutions is 0.3 * 0.13 = 0.04. The significance levels are indicated as: *p < 0.05; **p < 0.001; ***p < 0.0001. For more information about mediation analyses using structural equation models, see research methods essay by Ballen and Salehi (2021).

DISCUSSION

We set out to test the importance of institutional context on student experiences in biology courses using data generated collaboratively through a research coordination network (NSF DBI-1919462) from seven institutions across three institution types: community colleges, inclusive institutions, and selective institutions.

We show some patterns emerge across all institutions, regardless of institution type, such as the absence of performance gaps based on gender. However, we found that gaps in performance, test anxiety, and ESC varied across different institutions. Specifically, while we did not observe performance gaps based on race/ethnicity at the two community colleges, we did observe them at the inclusive and selective institutions in our sample. Second, we found a significant indirect effect of URM status mediated by test anxiety on exam scores at the inclusive institutions, but not at the community colleges or selective institutions. Third, we found overall that URM students at the selective institutions were more likely to report higher ESC values (but this did not in turn impact performance), while at community colleges, ESC positively affected exam scores for all students. Based on these results, we put forth a number of general and specific research recommendations.

First, we stress the importance of multiple perspectives in BER through programs that encourage research and collaborations across institution types. Other researchers have echoed the need for data-driven approaches to STEM education that span institutional contexts beyond selective institutions to address equity in STEM (Estrada *et al.*, 2016; Schinske *et al.*, 2017; Thompson *et al.*, 2020). For example, despite the important role of community colleges in educating students, only 3% of recent BER articles included either community college authors or a community college biology study context (Schinske *et al.*, 2017). And despite the unique role of community colleges in educating biology students from diverse backgrounds, Schinske *et al.* (2017) showed only 7% of community college BER articles explicitly studied equity and diversity. Our results support the notion that institutional context impacts barriers to student performance and interventions to support student success will likely vary in their effectiveness based on the student population. Beyond the coarse categories of institution type, we found individual institutions and classrooms also differ from one another. Specifically, we found that URM students underperformed relative to non-URM students on exams at both the inclusive and selective institutions, but not at the community colleges; however, individual courses showed variation within institution types, with the majority of courses showing no significant gap between groups. This supports the idea that factors other than precollege background, academic preparation, and admission criteria explain a large share of performance outcomes (Massey *et al.* 2003; Espenshade and Radford 2009; Salehi *et al.* 2019). Future research on factors that positively influence performance will clarify similarities across these equitable classrooms.

Second, social psychological interventions have been shown to vary in effectiveness for different students, and our results highlight the need for research into targeted interventions to support inclusion in STEM (Schwartz *et al.*, 2016). For example, we found that test anxiety was negatively associated with exam performance across institutions and mediated performance gaps for URMs at inclusive institutions. Future research would benefit from a focus on the impacts of test anxiety on performance of students with underrepresented identities in biology specifically. One study showed that lower measures of incoming preparation were associated with higher anxiety scores, which were associated with lower exam scores (Salehi *et al.*, 2020). While it is unclear why we only observed this mediating effect of anxiety at inclusive institutions, our results suggest that strategies instructors can use to mitigate test anxiety are likely to benefit all students, and especially URM students at some institutions. These include lowering the stakes of exams, expressive writing, or two-stage exams (Ergene, 2003; Ballen *et al.*, 2017a; Rempel *et al.*, 2021). A meta- analysis of 56 interventions that target student behavior ($N = 2482$) revealed that approaches such as skill-focused strategies (e.g., test-taking skills training), behavioral strategies (e.g., relaxation training), or cognitive strategies (e.g., techniques to change negative thought patterns) are effective at reducing test anxiety. However, less work has brought these strategies into biology or STEM classrooms to test their impacts on performance outcomes. Additionally, the impacts of test anxiety on performance of students with underrepresented identities in biology has not been studied as extensively, to our knowledge.

Conversely, given our varying findings about the effects of ESC, researchers and faculty should be intentional in their implementation of stereotype threat interventions in their institutions. While past work has demonstrated that, for students who have high ESC, the threat or fear of conforming to a stereotype negatively impacts their exam performance (McFarland *et al.*, 2003; Massey and Fischer, 2005), others have suggested that this phenomenon may not be as universal as previously thought (Cromley *et al.*, 2013).

Our results showed that overall, self-reported ESC did not negatively impact academic performance for URM students, and its mediating effects varied across institutions. This suggests stereotype threat did not impact performance of URM students across contexts. In fact, at community colleges, while URM status did not predict student ESC value, the ESC value positively predicted exam score (though the effect size was small). In other words, for all students, regardless of their minority status, self-reported ESC was associated with higher exam scores. We may observe that one's race/ethnicity does not undermine performance outcomes due to reduced negative stereotypes about race/ethnicity in more-inclusive contexts, such as community colleges. Future research would profit from an explicit focus on community colleges to understand this outcome and what other institutions can learn from equitable teaching practices that are effective at these institutions.

Fostering an inclusive environment serves to reduce the potential effects of stereotype threat (Steele, 2012) and is a leading explanation for positive performance outcomes among minoritized groups (Martin *et al.*, 2017). Murphy and Zirkel (2015) found that students in higher education anticipated a higher sense of belonging in majors where they perceived their group to be represented and that self-reported feelings of belonging in the first weeks of college predicted second-semester grades among students of color, but not White college students. We found that URM status was correlated with higher ESC at the selective institutions in our sample, supporting previous work that stresses the importance of promoting sense of belonging and social fit on campuses or institutions (Massey *et al.*, 2003; Martin *et al.*, 2017). The selective institutions in our sample include the lowest percentage of URM students (Table 2). These low numbers may lead students to feel marginalized and isolated (Hurtado and Ruiz, 2012; Patton and Croom, 2017), experience hostile learning environments (Kelly *et al.*, 2017), and possess few support networks (Palmer *et al.*, 2011). In these predominantly White spaces, we must center interventions that call upon instructors and institutions, rather than the students, to act. Previous research describes many ways to promote inclusion in the classroom, such as employing equitable teaching strategies (Tanner, 2013) and active learning (Eddy and Hogan 2014; Ballen *et al.* 2017b; Casper *et al.* 2019; Theobald *et al.* 2019), learning names (Cooper *et al.*, 2017), using humor (Cooper *et al.*, 2018), including positive role models for all students as examples in course content (Wood *et al.*, 2020; Yonas *et al.*, 2020), conveying that diversity is valued (Purdie-Vaughns *et al.*, 2008), and/or encouraging students to reflect on core personal values (Cohen *et al.*, 2006; Martens *et al.*, 2006; Miyake *et al.*, 2010).

Future research would profit from an understanding of how stereotypes specific to different underrepresented identities emerge in undergraduate settings, how they impact performance outcomes, and how students navigate these situations. Neal-Jackson (2020) conducted a qualitative analysis of Black women at a prestigious predominantly White institution and found that they encountered gendered racial stereotyping that negatively impacted their relationship to the academic community and their prospects for success. This, along with other work (Cho *et al.*, 2013; Morton and Parsons, 2018), also highlights the complexity associated with the multidimensionality of identity and the intersecting nature of those dimensions. We encourage researchers to continue exploring these ideas to determine what strategies increase academic performance and reduce threat for all students. Future work at community colleges and inclusive institutions will help us better understand how to alleviate ESC and inform interventions at selective institutions, where students from marginalized backgrounds are less likely to perceive social fit and more likely to experience stereotype threat.

Limitations

The work presented here is an initial step toward broadening the scope of discipline-based education research across different institutional contexts, though we faced a number of limitations that should be addressed in future work. We acknowledge that our approach here of grouping institutions by broad category of community college, inclusive, and selective suggests students share similar experiences with other students within the same broad institution type rather than across other factors, and this assumption is a limitation of our research. We do not expect that, for example, test anxiety fully mediates performance gaps across *all* inclusive 4 year-institutions in the United States. Our current sample includes 12 courses in two community colleges, six courses in three inclusive institutions, and 15 courses in three selective institutions. We administered surveys at a single time point, which might not be indicative of levels across the entire semester. Also, other social psychological measures (e.g., sense of belonging; Lewis *et al.*, 2016; Rainey *et al.*, 2018; Binning *et al.*, 2019) and their contribution to student performance should be explored.

Additionally, we did not have access to proxies for incoming preparation (e.g., ACT/Scholastic Aptitude Test [SAT]) from most institutions and were therefore unable to control for this factor in our analyses. Previous research demonstrates that, from the outset, capable but less academically prepared students are not provided with an equal opportunity to excel due to incoming preparation gaps and assessment norms in introductory science classes, resulting in the appearance of "performance gaps" in introductory classes, as we show here (Salehi *et al.*, 2019b). However, when incoming preparation is accounted for, these apparent gaps disappear. While we were not able to control for this, we note that student transcripts, applications to graduate and professional schools, and the internalization of poor performance do not take incoming preparation into account either (Harris *et al.*, 2020). Future research may test how incoming preparation impacts demographic performance gaps across institution types and control for incoming preparation in introductory classes while examining the effect of social psychological factors. While the lack of these data is a limitation, we do not believe this impacts our main conclusion, which points to variation in outcomes across institutional contexts.

Furthermore, one of our analyses included social psychological measures that we obtained from survey data, which represented a subset of all students. This subset may not be representative of the overall sample, as suggested by the URM performance gap in the whole sample as opposed to the gap in the subsample (no URM performance gaps for community colleges in the whole sample, as opposed to the 0.2 SD URM performance gap for community colleges in the subsample; Figure 1). Thus, our analyses focusing on the impacts of these factors on performance may not be representative of the larger pool of students across these institutional contexts. Therefore, the observed differences here in student experiences and academic performance outcomes warrant further study to examine to what extent these differences are results of differences in institutional contexts.

CONCLUSIONS

We call for future research to 1) not exclusively focus on selective institutions and instead include data across different types of institutions; 2) systematically examine characteristics of institutional or classroom contexts that help or hinder equity in undergraduate education; and 3) design targeted psychological interventions and/or effective teaching strategies that fit the characteristics of institutional or classroom contexts. We found some similar trends across institution types, as well as substantial differences with respect to the size of demographic performance gaps and the relationship between social psychological factors and performance. Ongoing research programs and existing evaluative tools designed to assess learning gains and social psychological factors lay a solid foundational groundwork for testing hypotheses.

ACKNOWLEDGMENTS

The institutions where these data were collected are located on the traditional lands of Indigenous Peoples, past and present, and we acknowledge and honor with gratitude the land itself and the people who have stewarded it through the generations. We are grateful to Auburn University's Biology Education Research group for constructive feedback, including Abby

Beatty, Emily Driessen, Sara Odom, and Todd Lamb. This work is supported by NSF DBI-1919462 awarded to S.C., C.J.B., and S.F.; DUE-2011995 and DUE-2120934 awarded to C.J.B. The research reported was approved or determined to be exempt from review by all local institutional review boards.

REFERENCES

Ali, M. S., & Mohsin, M. N. (2013). Relationship of test anxiety with students' achievement in science. *International Journal of Educational Science and Research*, *3*(1), 99–106.

American Association of Community Colleges. (2021). *Fast facts*. Retrieved May 20, 2021, from www.aacc.nche.edu/wp-content/uploads/2021/03/AACC_2021_FastFacts.pdf

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

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®)*. Washington, DC: American Psychiatric Association Publishing.

Appel, M., & Kronberger, N. (2012). Stereotypes and the achievement gap: Stereotype threat prior to test taking. *Educational Psychology Review*, *24*(4), 609–635.

Ballen, C. J., & Salehi, S. (2021). Mediation analysis in discipline-based education research using structural equation modeling: beyond "what works" to understand how it works, and for whom. *Journal of microbiology & biology education*, *22*(2), e00108-21.

Ballen, C. J., Salehi, S., & Cotner, S. (2017a). Exams disadvantage women in introductory biology. *PLoS ONE*, *12*(10), e0186419.

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

Binning, K. R., Cook, J. E., Purdie-Greenaway, V., Garcia, J., Chen, S., Apfel, N., ... & Cohen, G. L. (2019). Bolstering trust and reducing discipline incidents at a diverse middle school: How self-affirmation affects behavioral conduct during the transition to adolescence. *Journal of school psychology*, 75, 74–88.

Boucher, K. L., Rydell, R. J., Van Loo, K. J., & Rydell, M. T. (2012). Reducing stereotype threat in order to facilitate learning. *European Journal of Social Psychology*, *42*(2), 174–179.

Casper, A. M., Eddy, S. L., & Freeman, S. (2019). True grit: Passion and persistence make an innovative course design work. *PLoS Biology*, *17*(7), e3000359.

Chang, M. J., Cerna, O., Han, J., & Saenz, V. (2008). The contradictory roles of institutional status in retaining underrepresented minorities in biomedical and behavioral science majors. *Review of Higher Education*, *31*(4), 433–464.

Chapell, M. S., Blanding, Z. B., Silverstein, M. E., Takahashi, M., Newman, B., Gubi, A., & McCann, N. (2005). Test anxiety and academic performance in undergraduate and graduate students. *Journal of Educational Psychology*, *97*(2), 268.

Cho, S., Crenshaw, K. W., & McCall, L. (2013). Toward a field of intersectionality studies: Theory, applications, and praxis. *Signs: Journal of Women in Culture and Society*, *38*(4), 785–810.

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

Cooper, K. M., Haney, B., Krieg, A., & Brownell, S. E. (2017). What's in a name? The importance of students perceiving that an instructor knows their names in a high-enrollment biology classroom. *CBE—Life Sciences Education*, *16*(1), ar8.

Cooper, K. M., Hendrix, T., Stephens, M. D., Cala, J. M., Mahrer, K., Krieg, A., ... & Brownell, S. E. (2018). To be funny or not to be funny: Gender differences in student perceptions of instructor humor in college science courses. *PLoS ONE*, *13*(8), e0201258.

Cotner, S., Jeno, L. M., Walker, J. D., Jørgensen, C., & Vandvik, V. (2020). Gender gaps in the performance of Norwegian biology students: The roles of test anxiety and science confidence. *International Journal of STEM Education*, *7*(1), 1–10.

Crisp, G., Nora, A., & Taggart, A. (2009). Student characteristics, pre-college, college, and environmental factors as predictors of majoring in and earning a STEM degree: An analysis of students attending a Hispanic serving institution. *American Educational Research Journal*, *46*(4), 924–942.

Cromley, J. G., Perez, T., Wills, T. W., Tanaka, J. C., Horvat, E. M. N., & Agbenyega, E. T. B. (2013). Changes in race and sex stereotype threat among diverse STEM students: Relation to grades and retention in the majors. *Contemporary Educational Psychology*, *38*(3), 247–258. https://doi.org/10.1016/j.cedpsych.2013.04.003

DordiNejad, F. G., Hakimi, H., Ashouri, M., Dehghani, M., Zeinali, Z., Daghighi, M. S., & Bahrami, N. (2011). On the relationship between test anxiety and academic performance. *Procedia—Social and Behavioral Sciences*, *15*, 3774–3778.

Eddy, S. L., & Hogan, K. A. (2014). Getting under the hood: How and for whom does increasing course structure work? *CBE—Life Sciences Education*, *13*(3), 453–468.

Ergene, T. (2003). Effective interventions on test anxiety reduction: A meta-analysis. *School Psychology International*, *24*(3), 313–328. https://doi.org/10.1177/01430343030243004

Espenshade, T. J., & Radford, A. W. (2009). *No longer separate, not yet equal*. Princeton, NJ: Princeton University Press.

Estrada, M., Burnett, M., Campbell, A. G., Campbell, P. B., Denetclaw, W. F., Gutiérrez, C. G., ... & Zavala, M. E. (2016). Improving underrepresented minority student persistence in stem. *CBE—Life Sciences Education*, *15*(3), 1–10. https://doi.org/10.1187/cbe.16-01-0038

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 USA*, *111*(23), 8410–8415.

Good, C., Aronson, J., & Inzlicht, M. (2003). Improving adolescents' standardized test performance: An intervention to reduce the effects of stereotype threat. *Journal of Applied Developmental Psychology*, *24*(6), 645–662.

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. https://doi.org/10.1126/science.1204820

Harris, R. B., Grunspan, D. Z., Pelch, M. A., Fernandes, G., Ramirez, G., & Freeman, S. (2019). Can test anxiety interventions alleviate a gender gap in an undergraduate STEM course? *CBE—Life Sciences Education*, *18*(3), ar35.

Harris, R. B., Mack, M. R., Bryant, J., Theobald, E. J., & Freeman, S. (2020). Reducing achievement gaps in undergraduate general chemistry could lift underrepresented students into a "hyperpersistent zone." *Science Advances*, *6*(24), 1–9. https://doi.org/10.1126/sciadv.aaz5687

Hong, E., & Karstensson, L. (2002). Antecedents of state test anxiety. *Contemporary Educational Psychology*, *27*(2), 348–367.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hurtado, S., Eagan, M. K., Tran, M. C., Newman, C. B., Chang, M. J., & Velasco, P. (2011). "We do science here": Underrepresented students' interactions with faculty in different college contexts. *Journal of Social Issues*, *67*(3), 553–579.

Hurtado, S., Newman, C. B., Tran, M. C., & Chang, M. J. (2010). Improving the rate of success for underrepresented racial minorities in STEM fields: Insights from a national project. *New Directions for Institutional Research*, *2010*(148), 5–15.

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.

Hurtado, S., & Ruiz Alvarado, A. (2015). Discrimination and bias, underrepresentation, and sense of belonging on campus.

Kellow, J. T., & Jones, B. D. (2008). The effects of stereotypes on the achievement gap: Reexamining the academic performance of African American high school students. *Journal of Black Psychology*, *34*(1), 94–120. https://doi.org/10.1177/0095798407310537

Kelly, B. T., Segoshi, M., Adams, L., & Raines, A. (2017). Experiences of Black alumnae from PWIs: Did they thrive? *NASPA Journal About Women in Higher Education*, *10*(2), 167–185.

Knekta, E., Runyon, C., & Eddy, S. (2018). One size doesn't fit all: Using factor analysis to gather validity evidence when using surveys in your research. *CBE—Life Sciences Education*, *18*(1), 1–17. https://doi.org/10.1187/cbe.18-04-0064

Koester, B. P., Grom, G., & McKay, T. A. (2016). Patterns of gendered performance difference in introductory STEM courses. Retrieved from arXiv: 1608.07565.

Koo, K. K. (2021). Am I welcome here? Campus climate and psychological well-being among students of color. *Journal of Student Affairs Research and Practice*, *58*(2), 196–213.

Lewis, K. L., Stout, J. G., Pollock, S. J., Finkelstein, N. D., & Ito, T. A. (2016). Fitting in or opting out: A review of key social-psychological factors influencing a sense of belonging for women in physics. *Physical Review Physics Education Research*, *12*(2), 020110.

Martens, A., Johns, M., Greenberg, J., & Schimel, J. (2006). Combating stereotype threat: The effect of self-affirmation on women's intellectual performance. *Journal of Experimental Social Psychology*, *42*(2), 236–243.

Martin, N. D., Spenner, K. I., & Mustillo, S. A. (2017). A test of leading explanations for the college racial-ethnic achievement gap: Evidence from a longitudinal case study. *Research in Higher Education*, *58*(6), 617–645.

Massey, D. S., Charles, C. Z., Fischer, M. J., & Lundy, G. (2003). *The source of the river: The social origins of freshmen at America's selective colleges and universities* (Vol. *36*). Princeton, NJ: Princeton University Press.

Massey, D. S., & Fischer, M. J. (2005). Stereotype threat and academic performance: New findings from a racially diverse sample of college freshmen. *Du Bois Review*, *2*(1), 45–67. https://doi.org/10.1017/S1742058X05050058

Matthews, G., Hillyard, E. J., & Campbell, S. E. (1999). Metacognition and maladaptive coping as components of test anxiety. *Clinical Psychology & Psychotherapy*, *6*(2), 111–125. https://doi.org/10.1002/(sici)1099-0879(199905)6:2<111::Aid-cpp192>3.3.co;2-w

McFarland, L. A., Lev-Arey, D. M., & ZIegert, J. C. (2003). An examination of stereotype threat in a motivational context. *Human Performance*, *16*(3), 181–205. https://doi.org/10.1207/S15327043HUP1603

Mervis, J. (2011). Weed-out courses hamper diversity. *Science*, *334*(6061), 1333.

Miyake, A., Kost-Smith, L. E., Finkelstein, N. D., Pollock, S. J., Cohen, G. L., & Ito, T. A. (2010). Reducing the gender achievement gap in college science: A classroom study of values affirmation. *Science*, *330*(6008), 1234–1237.

Morton, T. R., & Parsons, E. C. (2018). # BlackGirlMagic: The identity conceptualization of Black women in undergraduate STEM education. *Science Education*, *102*(6), 1363–1393.

Murphy, M. C., & Zirkel, S. (2015). Race and belonging in school: How anticipated and experienced belonging affect choice, persistence, and performance. *Teachers College Record*, *117*(12), 1–40.

National Academy of Sciences, National Academy of Engineering, and Institute of Medicine. (2011). *Expanding underrepresented minority participation: America's science and technology talent at the crossroads*. Washington, DC: National Academies Press.

National Institutes of Health. (2015). *Racial and ethnic categories and definitions for NIH diversity programs and for other reporting purposes*. U.S. Department of Health and Human Services.

National Science Foundation. (2019). *Women, minorities, and persons with disabilities in science and engineering. Proceedings from NCSES, NSF 19-304. March 8, 2019*.

Neal-Jackson, A. (2020). "Well, what did you expect?": Black women facing stereotype threat in collaborative academic spaces at a predominantly white institution. *Journal of College Student Development*, *61*(3), 317–332.

Offerdahl, E. G., Balser, T., Dirks, C., Miller, K., Momsen, J. L., Montplaisir, L., … & White, B. (2011). Society for the Advancement of Biology Education Research (SABER). *CBE—Life Sciences Education*, *10*(1), 11–13.

Osborne, J. W., & Walker, C. (2006). Stereotype threat, identification with academics, and withdrawal from school: Why the most successful students of colour might be most likely to withdraw. *Educational Psychology*, *26*(4), 563–577. https://doi.org/10.1080/01443410500342518

Page, K. R., Castillo-Page, L., Poll-Hunter, N., Garrison, G., & Wright, S. M. (2013). Assessing the evolving definition of underrepresented minority and its application in academic medicine. *Academic Medicine*, *88*(1), 67–72.

Palmer, R. T., Maramba, D. C., & Holmes, S. L. (2011). A contemporary examination of factors promoting the academic success of minority students at a predominantly white university. *Journal of College Student Retention: Research, Theory & Practice*, *13*(3), 329–349.

Park, J. J., Kim, Y. K., Salazar, C., & Eagan, M. K. (2020). Racial discrimination and student–faculty interaction in STEM: Probing the mechanisms influencing inequality. *Journal of Diversity in Higher Education*. https://doi.org/10.1037/dhe0000224

Patton, L. D., & Croom, N. N. (2017). *Critical perspectives on Black women and college success*. England, UK: Taylor & Francis.

Picho, K., & Brown, S. W. (2011). Can stereotype threat be measured? A validation of the Social Identities and Attitudes Scale (SIAS). *Journal of Advanced Academics*, *22*(3), 374–411.

Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1993). Reliability and predictive validity of the Motivated Strategies for Learning Questionnaire (MSLQ). *Educational and Psychological Measurement*, *53*(3), 801–813.

President's Council of Advisors on Science and Technology. (2012). *Engage to excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics*. Washington, DC: U.S. Government Office of Science and Technology.

Purdie-Vaughns, V., Steele, C. M., Davies, P. G., Ditlmann, R., & Crosby, J. R. (2008). Social identity contingencies: How diversity cues signal threat or safety for African Americans in mainstream institutions. *Journal of Personality and Social Psychology*, *94*(4), 615.

Rainey, K., Dancy, M., Mickelson, R., Stearns, E., & Moller, S. (2018). Race and gender differences in how sense of belonging influences decisions to major in STEM. *International journal of STEM education*, *5*(1), 1–14.

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rempel, B. P., Dirks, M. B., & McGinitie, E. G. (2021). Two-stage testing reduces student-perceived exam anxiety in introductory chemistry. *Journal of Chemical Education*, *98*(8), 2527–2535.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of statistical software*, *48*(2), 1–36.

Rydell, R. J., Rydell, M. T., & Boucher, K. L. (2010). The effect of negative performance stereotypes on learning. *Journal of Personality and Social Psychology*, *99*(6), 883.

Salehi, S., Burkholder, E., Lepage, G. P., Pollock, S., & Wieman, C. (2019a). Demographic gaps or preparation gaps?: The large impact of incoming preparation on performance of students in introductory physics. *Physical Review Physics Education Research*, *15*(2), 20114.

Salehi, S., Cotner, S., Azarin, S. M., Carlson, E. E., Driessen, M., Ferry, V. E., … & Ballen, C. J. (2019b). Gender performance gaps across different assessment methods and the underlying mechanisms: The case of incoming preparation and test anxiety. *Frontiers in Education*, *4*, 107. Retrieved from www.frontiersin.org/article/10.3389/feduc.2019.00107

Salehi, S., Cotner, S., & Ballen, C. J. (2020). Variation in incoming academic preparation: Consequences for minority and first-generation students. *Frontiers in Education*, *5*. https://doi.org/10.3389/feduc.2020.552364

Sarason, I. G. (1961). Test anxiety and the intellectual performance of college students. *Journal of Educational Psychology*, *52*(4), 201–206.

Schinske, J. N., Balke, V. L., Bangera, M. G., Bonney, K. M., Brownell, S. E., Carter, R. S., … & Corwin, L. A. (2017). Broadening participation in biology education research: Engaging community college students and faculty. *CBE—Life Sciences Education*, *16*(2), mr1. https://doi.org/10.1187/cbe.16-10-0289

Schwartz, D. L., Cheng, K. M., Salehi, S., & Wieman, C. (2016). The half empty question for socio-cognitive interventions. *Journal of Educational Psychology*, *108*(3), 397–404. https://doi.org/10.1037/edu0000122

Seymour, E., & Hewitt, N. M. (1997). *Talking about leaving: Why undergraduates leave the sciences*. Boulder, CO: Westview Press.

Seymour, E., & Hunter, A. B. (2019). *Talking about leaving revisited. Talking About Leaving Revisited: Persistence, Relocation, and Loss in Undergraduate STEM Education*. Switzerland AG*:* Springer Nature*.*

Shapiro, A. L. (2014). Test anxiety among nursing students: A systematic review. *Teaching and Learning in Nursing*, *9*(4), 193–202.

Smit, R. (2012). Towards a clearer understanding of student disadvantage in higher education: Problematising deficit thinking. *Higher Education Research & Development*, *31*(3), 369–380.

Spencer, S. J., Logel, C., & Davies, P. G. (2016). Stereotype threat. *Annual Review of Psychology*, *67*, 415–437.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of experimental social psychology*, *35*(1), 4–28.

Stains, M., Harshman, J., Barker, M. K., Chasteen, S. V., Cole, R., DeChenne-Peters, S. E., ... & Young, A. M. (2018). Anatomy of STEM teaching in North American universities. *Science*, *359*(6383), 1468–1470.

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. Retrieved October 26, 2021, from www.ncbi.nlm.nih.gov/pubmed/7473032

Taasoobshirazi, G., Puckett, C., & Marchand, G. (2019). Stereotype threat and gender differences in biology. *International Journal of Science and Mathematics Education*, *17*(7), 1267–1282.

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

Taylor, V. J., & Walton, G. M. (2011). Stereotype threat undermines academic learning. *Personality and Social Psychology Bulletin*, *37*(8), 1055–1067.

Theobald, E. J., Aikens, M., Eddy, S., & Jordt, H. (2019). Beyond linear regression: A reference for analyzing common data types in discipline based education research. *Physical Review Physics Education Research*, *15*(2), 20110.

Theobald, E. J., Hill, M. J., Tran, E., Agrawal, S., Arroyo, E. N., Behling, S., ... & Dunster, G. (2020). Active learning narrows achievement gaps for underrepresented students in undergraduate science, technology, engineering, and math. *Proceedings of the National Academy of Sciences USA*, *117*(12), 6476–6483.

Thompson, S., Berk, S., Hall, C., Creech, C., Harshman, J., Garcia Ojeda, M., ... & Ballen, C. J. (2020). A call for data-driven networks to address equity in the context of undergraduate biology. *CBE—Life Sciences Education* (*in-review*).

Topp, R. (1989). Effect of relaxation or exercise on undergraduates' test anxiety. *Perceptual and Motor Skills*, *69*, 35–41.

von der Embse, N., Jester, D., Roy, D., & Post, J. (2018). Test anxiety effects, predictors, and correlates: A 30-year meta-analytic review. *Journal of Affective Disorders*, *227*, 483–493.

Wang, X. (2013). Modeling entrance into STEM fields of study among students beginning at community colleges and four-year institutions. *Research in Higher Education*, *54*(6), 664–692.

Winkle-Wagner, R., & McCoy, D. L. (2018). Feeling like an "alien" or "family"? Comparing students and faculty experiences of diversity in STEM disciplines at a PWI and an HBCU. *Race Ethnicity and Education*, *21*(5), 593–606.

Wood, S., Henning, J. A., Chen, L., McKibben, T., Smith, M. L., Weber, M., ... & Ballen, C. J. (2020). A scientist like me: Demographic analysis of biology textbooks reveals both progress and long-term lags. *Proceedings of the Royal Society of London B*, *287*(1929), 20200877.

Yonas, A., Sleeth, M., & Cotner, S. (2020). In a "Scientist Spotlight" intervention, diverse student identities matter. *Journal of Microbiology & Biology Education*, *21*(1), 21.1.15

Zhao, Y. (2016). From deficiency to strength: Shifting the mindset about education inequality. *Journal of Social Issues*, *72*(4), 720–739.