Title

Evaluating a National Biomedical Training Program Using QuantCrit: Revealing Disparities in Research Self-efficacy for Women of Color Undergraduates.

Journal

CBE—Life Sciences Education, 23(4)

ISSN

1931-7913

Authors

Srinivasan, Jayashri
Cobian, Krystle P
Maccalla, Nicole MG
et al.

Publication Date

2024-12-01

DOI

10.1187/cbe.24-02-0047

Peer reviewed

Evaluating a National Biomedical Training Program Using QuantCrit: Revealing Disparities in Research Self-efficacy for Women of Color Undergraduates

Jayashri Srinivasan,[†]* Krystle P. Cobian,[‡] Nicole M. G. Maccalla,[§] and Christina A. Christie[‖]

[†]Associate Project Scientist, Coordination and Evaluation Center, University of California, Los Angeles (UCLA), Los Angeles, CA 90025; [‡]Investigator, UCLA Coordination and Evaluation Center Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA 90025; [§]Co-Director of Evaluation, Coordination and Evaluation Center, University of California, Los Angeles, Los Angeles, CA 90025; [‖]Wasserman Dean and Professor of Social Research Methodology, School of Education and Information Studies, University of California, Los Angeles, CA 90025

ABSTRACT

Program evaluation for interventions aimed at enhancing diversity can fall short when the evaluation unintentionally reifies the exclusion of multiple marginalized student experiences. The present study presents a Quantitative Critical Race Theory (QuantCrit) approach to program evaluation to understand outcomes for Women of Color undergraduates involved in a national biomedical training program called the Building Infrastructure Leading to Diversity (BUILD) initiative. Using longitudinal data, we examined the impact of participation in the BUILD Scholars programs and BUILD-developed novel biomedical curriculum on undergraduate's research self-efficacy. Employing propensity score matching and multiple regression models, we found that Black women who participated in the BUILD scholars program reported higher research self-efficacy, whereas Latine and White undergraduate BUILD scholars had lower research self-efficacy. Additionally, Latine women who participated in novel biomedical curricula reported significantly lower research self-efficacy. We contend that disaggregated and intersectional analyses of subpopulations are necessary for improving understanding of program interventions and identifying areas where systems of exclusion may continue to harm students from minoritized backgrounds. We provide recommendations for future quantitative program evaluation practices and research in science, technology, engineering, mathematics, and medicine (STEMM) equity efforts.

INTRODUCTION AND PURPOSE

Program evaluation of science, technology, engineering, mathematics, and medicine (STEMM) intervention programs are a growing priority for STEMM education and workforce development initiatives (Mertens and Hopson, 2006). National efforts aimed at enhancing the scientific workforce from agencies such as the National Science Foundation (NSF) and National Institutes of Health (NIH) have recently shifted toward a greater emphasis on social science research to study "what works," "for whom," "why," and "how" regarding STEMM intervention programs (Tsui, 2007; Gibbs *et al.*, 2022; Maccalla *et al.*, 2022). To this end, funding agencies are driving efforts to enhance the evaluation of STEMM equity initiatives aimed at "enhancing

Terrell Morton, *Monitoring Editor*

Submitted Feb 7, 2024; Revised Sep 5, 2024; Accepted Sep 16, 2024

CBE Life Sci Educ December 1, 2024 23:ar54

DOI:10.1187/cbe.24-02-0047

Conflicts of interest: The authors declare no conflicts of interest.

*Address correspondence to: Jayashri Srinivasan (Jsrini@ucla.edu).

diversity," "broadening participation," and/or "ending structural racism" (Collins *et al.*, 2021).

In 2014, the NIH launched Enhancing the Diversity of the NIH-Funded Workforce, also known as the Diversity Program Consortium (DPC), a national collaborative consisting of the Building Infrastructure Leading to Diversity (BUILD) initiative and the Coordination and Evaluation Center (CEC), designed to develop, implement, assess, and disseminate effective and innovative approaches to mentoring and research training for students, faculty, and institutions (National Institute of General Medical Sciences (NIGMS, n.d.)). This study focuses on the BUILD initiative, which consists of 10 awardee sites (seven minority-serving institutions) at 11 universities across the United States. By implementing and evaluating effective approaches to improve training and mentoring across all the awardee sites, the program aims to have a significant impact on the participation and persistence of individuals from diverse backgrounds throughout the biomedical research pathway.

There are several epistemological and methodological challenges to evaluating programs where enhancing diversity is an aim. First, STEMM program evaluation involves interdisciplinary collaboration among scientists, social scientists, and program evaluators, with each field having various norms, values, and training methods that place emphasis on various epistemologies and methodologies for knowledge development (Mertens, 1999; Katzenmeyer and Lawrenz, 2006). Second, traditional program evaluation tends to focus on "what works" for most participants, which can decenter the experiences of participants from minoritized groups. For example, normative evaluation consists of assessing one's performance compared with other individuals. In practice, this can often prioritize using advanced statistical procedures to calculate a treatment effect (aggregated) over the nuanced analyses of treatment effects on marginalized subgroups (disaggregated). Another challenge is epistemology: the positivist paradigm prevalent in science has been criticized for assuming neutrality and objectivity, contributing to faulty interpretations of analyses without deep consideration for how power can shape a study's questions, design, and interpretation of results. Resultingly, program evaluators face pressure to report results regarding the extent to which an intervention worked (Morris and Clark, 2013) through measures and interpretations deemed objective and impartial, leading to a hasty interpretation that a program works the same for all participants in all contexts. Lastly, program evaluation might emphasize positive outcomes rather than emphasizing how racism, sexism, or other axes of oppression shape study results and efforts toward social change (Stage, 2007; Covarrubias and Velez, 2013; Sablan, 2019; Tabron and Thomas, 2023).

Without a critical lens applied to all steps of the study (Suzuki *et al.*, 2021), an evaluation can fall short of identifying key findings that reveal structural challenges and provide a nuanced understanding of the differential outcomes and how sociopolitical contexts shape such outcomes. Thus, we propose program evaluation activities that leverage critical epistemologies and methods to advance research findings and develop methods of "knowing" that might be more accurate and better aligned with improving the outcomes for marginalized groups (Mertens, 1999), and thus more likely to truly enhance participation and persistence of individuals from underrepresented groups. We provide an example of a critical paradigm-driven approach (Stage and Wells, 2014; Garcia *et al.*, 2018; Maass *et al.*, 2018) to evaluate STEMM intervention programs and describe instances throughout this study to illustrate how a critical evaluation approach is distinct from typical normative and positivist approaches to program evaluation.

This evaluation examined the BUILD initiative, a national biomedical program within the larger NIH-funded DPC. We employed quantitative, longitudinal student survey data from the Enhance Diversity Study, a nationwide evaluation of the 10 NIH-funded BUILD sites, to examine changes in undergraduates' research self-efficacy (RSE), with a particular focus on Women of Color (WOC). Our rationale for focusing on WOC in STEMM is that they are often decentered when analyses focus on aggregate treatment effects for race and gender separately (Bowleg, 2008) or only named implicitly rather than "making raced women visible" (Miles *et al.*, 2022, p.230). While their representation in biomedical disciplines has increased over the past decades, increased participation does not mean equal outcomes or better experiences while on a STEMM training trajectory (McGee and Robinson, 2020). Prior research demonstrated a strong positive effect of BUILD participation on RSE (Cobian *et al.*, 2021; Crespi and Cobian, 2022). This study addresses an important and unanswered equity-focused evaluation question, "for whom does BUILD work?" (Maccalla *et al.*, 2022), and how racism and other systems of marginalization may be operating and shaping outcomes for WOC.

This study addresses the following evaluation questions:

1. Are there differences in RSE at college entry (i.e., at baseline) across Gender and Racial Ethnic groups for students at BUILD institutions?
2. Is there a differential effect of the BUILD Scholars program on RSE for undergraduate WOC biomedical majors?
3. Is there a differential effect of participating in BUILD novel curriculum for undergraduate WOC biomedical majors?

LITERATURE REVIEW AND THEORETICAL FRAMEWORK

WOC in STEMM

Despite efforts to broaden the diversity of groups historically underrepresented in science, WOC are still numerically underrepresented in several STEMM fields (Williams, 2014; McGee *et al.*, 2021). WOC remain almost invisible in fields such as computer science and engineering (Mack, Rankins, and Woodson, 2013; National Research Council, 2010). While their participation in biological sciences and health sciences has increased in the past three decades, in 2022 only 11.3% of all U.S. citizens and permanent resident doctoral students in science, engineering, and health were WOC, compared with White women (18.1%), Men of Color (21.2%), and White men (36.1%), indicating the gaps in participation by race and gender (National Center for Science and Engineering Statistics, 2022). Regarding their experiences along their STEM training and career paths, WOC in STEMM experience isolation (Ong *et al.*, 2011) and stereotype threat (Collins *et al.*, 2020). WOC continue to be underrepresented in R01 research grants (Ginther *et al.*, 2016); a study of the probability of being awarded an R01 grant from the NIH between 2000 and 2006, Asian and Black women PhDs were significantly less likely to

receive grant funding than White women and men (Ginther *et al.*, 2016). WOCs are further marginalized in their research impact and, thus, research career advancement (Kozlowski *et al.*, 2022). These disparities are critical on their own, but especially urgent to address considering that WOC are expected to become the majority of the U.S. population of women by the year 2060 (Catalyst, 2023; United States Census Bureau, 2017).

Scholarship on WOC shows how representation is shaped by power structures and norms that create additional barriers that WOC must face in comparison to White women or men of color. WOC experience raced- and gendered-interactions in White-men-normed STEMM environments, which can add additional stress. For example, in a study of STEMM doctoral students from diverse racial and ethnic backgrounds, the compound effect of stressors took a psychological toll that influenced their decisions to remain in STEMM fields (Wilkins-Yel *et al.*, 2022). Indeed, experiences of discrimination, racialized and gendered pressure to succeed, stereotype threat (Collins *et al.*, 2020), and isolation experienced by WOC in STEMM point to the utility of Intersectionality as a framework for understanding how power structures operate and uniquely shape both the psychological and material experiences and outcomes for WOC.

QuantCrit and Intersectionality as Key Frameworks

Derived from Critical Race Theory (CRT), QuantCrit is a methodological subfield that emerged from Critical Race Studies in Education to advance quantitative research for racial justice (Garcia *et al.*, 2018; Castillo and Babb, 2024). Where quantitative research is often described as positivist or postpositivist (i.e., assuming truth is objective), QuantCrit scholarship seeks to imagine and rectify assumptions of objectivity in quantitative research methods by employing a critical ontological and epistemological stance (*Garcia et al.*, 2018, 2023; Gillborn *et al.*, 2018; Sablan, 2019). There are five principles of QuantCrit: 1) the centrality of racism; 2) numbers are not neutral; 3) categories are not natural; 4) voice and insight (positing that data cannot 'speak for itself"), and critical analyses must be informed by experiential knowledge of marginalized groups; and 5) orientation toward social justice (Garcia *et al.*, 2018; Gillborn *et al.*, 2018), employed throughout the research process (Suzuki *et al.*, 2021).

Intersectionality is a framework concerned with how systems of oppression intertwine and influence individuals' life experiences and opportunities (Combahee River Collective, 2014; Crenshaw, 1991; Collins and Bilge, 2016). Considering STEMM interventions aimed at enhancing participation in the STEMM workforce, Intersectionality places increased attention on the barriers to science that WOC encounter rather than the individual attributes that predispose individual WOC to persist (Collins and Bilge, 2016). Moreover, an Intersectionality framework shifts the interpretation of a study to discuss how systems of power might be shaping the results. For example, López and colleagues (2018) utilized interaction terms to analyze outcomes at the intersections of race, gender, and class, which revealed differential outcomes in 6-year postsecondary college graduation rates. They interpreted their analysis as modeling social locations as a relational social status in society to reveal the complexity of educational disparities. Also employing intersectionality, Ovink *et al.* (2024) examined students' race/ethnicity, gender, and academic major to understand how one's social locations (and thus, relations to systems of oppression) were related to academic and social belonging. Both studies emphasize that their results should not be used to justify racial stratification but to uncover systematized exclusion and aspire toward social justice (Collins and Bilge, 2016).

A QuantCrit Approach to Program Evaluation

Program evaluation is a systematic inquiry used to ascertain the merit or worth of a program or intervention (Alkin and Vo, 2018). Distinct from traditional research, evaluation is bounded within a specific context, with the focus of inquiry being driven by clients, funders, and interested parties, with the explicit intention to use findings for decision-making (Alkin and Vo, 2018). Summative evaluation, which is outcome-focused, helps funders decide whether a program has successfully achieved its stated goals (i.e., "Did it work?") (Alkin and Vo, 2018). Impact evaluation is a term used to describe a methodological approach to determining whether the observed changes in desired outcomes can be attributed to the program or intervention being evaluated (Rogers, 2014).

One critique of impact evaluation is the tendency to mask nuanced experiences in favor of analyses that determine whether an intervention "works" or not. These longitudinal studies often favor (quasi)experimental designs, larger sample sizes, and statistical power but ignore the fact that averages can mask within-group differences and inequities. While impact evaluation can inform programmatic decisions, looking only at high-level outcomes can have potentially negative consequences for individuals who might not truly benefit, or benefit differently, from the intervention. While there is some guidance on implementing QuantCrit in research studies (Milner 2007; Suzuki *et al.*, 2021; Castillo and Gillborn, 2023), little exists on how to do so in a large-scale evaluation study. A longitudinal quantitative study design that looks at "for whom" and "under what contexts" advances evaluation and aligns with tenets of QuantCrit and Intersectionality, which seek to look beyond individuals' identities by examining power structures that might contribute to outcomes.

A QuantCrit approach to program evaluation differs because it employs a critical lens that guides the evaluation questions, analytic approaches, and interpretation of findings. Evaluators ask questions about themselves, each other, and the data, establishing a critical sensibility in the work (Milner, 2007; Castillo and Gillborn, 2023). Suzuki *et al.* (2021) described three "moments" of the research process and how researchers can apply QuantCrit principles within them. We employed these critical reflection moments and expand by proposing an additional "moment" that is unique to the program evaluation process.

In the first "moment" of developing research questions and variables (Suzuki *et al.*, 2021), it is important to frame guiding questions from an asset-based perspective, shifting the focus from the individuals in the study (i.e., WOC) to the system (i.e., BUILD program, institution, and/or broader biomedical research training culture in the United States) (Milner, 2007; Castillo and Gillborn, 2023). For example, the guiding evaluation questions for this study focus on the differential effects

of the BUILD program for WOC rather than disparities in RSE outcomes attributed to WOC in BUILD. Using guiding frameworks such as CRT and Intersectionality is an additional way to apply QuantCrit principles in this "moment."

In the second "moment" of determining the role of race in analytic models, we deliberated at great length about how to QuantCrit within a pre-existing evaluation project focused on equity-driven evaluation rather than QuantCrit. We addressed the reality of racism operating through decisions made about variable creation, modeling decisions, and statistical analyses that aimed to "deracialize" (Suzuki *et al.*, 2021). A QuantCrit approach requires awareness of and attention to issues, policies, and practices that have served to oppress certain groups in the United States, particularly in STEMM education, which guides us toward the choice of variables and how they might be included in our models.

Model selection is important in this second "moment" (Suzuki *et al.*, 2021; Castillo and Gillborn, 2023). Quantitative approaches that test the impact of a program on an outcome can inadvertently obscure the role of Intersectionality and thus promote and reinforce faulty assumptions about multiple marginalized groups (Pearson *et al.*, 2022). For example, in regression analyses, researchers may control for underrepresented identities of interest, such as race/ethnicity, gender, disability status, etc., to account for differences in social identities or experiences of individuals in the sample. While race and ethnicity, as well as gender, might technically be added as variables in a regression model, individual aspects of identity are at best reduced one by one into mutually exclusive dichotomous variables, with each comparison being made to a single reference group with power (e.g., Black vs. White; women vs. men), or at worst, and often with small sample sizes, into gross binary race groups (e.g., White vs. non-White) categories which can perpetuate racist narratives (Suzuki *et al.*, 2021). To address this challenge, a QuantCrit approach utilizes statistical methods from a critical standpoint that interrogates power structures. Instead of an interpretation that explains how a model controlled for race and gender, the interpretation of the model is about utilizing the variables and sample to statistically understand the extent to which racism and sexism might be revealed by differential outcomes. In other words, including a race variable in a model comes with the assumption that race is important to examine because racism is a persistent and pervasive part of society. A QuantCrit approach focused on Intersectionality must also allow for multiple and layered race and ethnicity group comparisons happening simultaneously, which may more closely mirror participants' self-identification and lived experience (i.e., developing models that account for one's race and gender simultaneously rather than separately).

QuantCrit aims to rectify how quantitative methods might be conducted and interpreted. To study whether BUILD activities differentially impacted WOC, we use the propensity score approach to create similar treatment/control groups. One of the challenges when studying the impact of an intervention includes working with small sample sizes (often problematic for sexual and gender minority variables and small or hard-to-reach populations), unbalanced treatment and control groups, and unequal variances between groups. These challenges shape what we can study regarding groups historically excluded from full participation and success in STEMM. Indeed, Gelman and Hill (2006) note:

> When treatment and control groups are *unbalanced*, the simple comparison of group averages, $\bar{y}_1 - \bar{y}_0$, is not generally a good estimate of the average treatment effect (ATE). Instead, some analysis must be performed to adjust for pretreatment differences between the groups. … When treatment and control groups do not completely overlap, the data are inherently limited in what they can tell us about treatment effects in the regions of nonoverlap. No amount of adjustment can create direct treatment/control comparisons, and one must either restrict inferences to the region of overlap or rely on a model to extrapolate outside this region. (p. 199).

Thus, balancing groups on multiple baseline covariates to create one propensity score helps create similar treatment and control groups to examine the effect of an intervention. It is worth noting that the selection of covariates should be chosen with care and deliberation in the process of critical inquiry, thereby acknowledging the engrained nature of racism in the fabric of society, even in the variables we may be controlling for (Castillo and Gillborn, 2023).

The third "moment" calls for interpreting results through a critical framework (Suzuki *et al.*, 2021). An impact evaluation from a positivist lens might faultily interpret the results of inferential models such that low-income "x" race/ethnicity students are more or less likely to score higher/lower on an outcome variable. However, a QuantCrit analysis in an evaluation would carefully contextualize the data and understand what structures, systems, policies, and practices are shaping the outcomes, not the other way around.

Overall, a QuantCrit approach to evaluation differs from a positivist quantitative study in its critical epistemological stance, focusing on anti-racism in its design, conceptualization of variables, analytical decisions, interpretation of results, and sharing these results with program facilitators and funders (Mertens, 1999; Castillo and Gillborn, 2023; Suzuki *et al.*, 2021). In the results section, we use the frameworks of QuantCrit and Intersectionality to make sense of the findings. The discussion and implications section details how program evaluation, with its prioritization of using findings for decision-making, uniquely lends itself to an additional fourth QuantCrit "moment."

The BUILD Initiative

BUILD programs implement and study various approaches to engage and retain students from diverse backgrounds in biomedical research National Institute of General Medical Sciences (NIGMS, n.d.). Eligibility for the BUILD award included having a student population with at least 25% Pell Grant recipients and fewer than $7.5 million in total NIH research grant funding (National Institute of General Medical Sciences (NIGMS, n.d.). The BUILD programs are situated within each awardee site's unique institutional context (Cobian *et al.*, 2024), which is situated within the broader context of postsecondary STEMM training. Each of these levels is part of the sociopolitical-cultural context where multiple systems of power operate and shape outcomes for individuals.

Considering the context, criteria for BUILD likely narrowed the pool of applicant institutions to those with a large proportion of Pell Grant recipients and historically little investment from NIH research grant funding. Often, these institutions serve an economically and racially diverse undergraduate population. Indeed, seven of the BUILD awardees are minority-serving institutions (Cobian *et al.*, 2024.

BUILD Scholars. While each BUILD award is uniquely implemented at each institution, in this study, students intensely involved in these programs are referred to as *BUILD Scholars*, that is, those who participated in these programs as Scholars and/or Associates. They often receive tuition support or a stipend, research training, Undergraduate Research Experiences (URE), academic and professional development opportunities, a learning community, and mentorship. Some BUILD-related activities are compulsory, while others are optional. Some are available only to Scholars and Associates, and others are open access (i.e., accessible to all undergraduate students). Over 2000 students have been enrolled in BUILD as Scholars and/or Associates.

BUILD Novel Biomedical Curriculum. One such open-access opportunity takes place through *Novel Curriculum*, which includes over 125 newly designed or revamped courses across the 10 BUILD sites geared to increase student engagement in biomedical sciences. Originally BUILD-sponsored and now institutionalized, these courses serve to increase and sustain the reach and impact of programming. The BUILD sites proposed and implemented a variety of curricular changes in STEMM departments (Goodwin *et al.*, 2021; Guerard and Hayes, 2018; Urizar and Miller, 2022). For example, the BUILD program based at the University of Texas at El Paso (UTEP) developed the Freshman Year Research Intensive Sequence (FYRIS), a multicourse sequence that integrated URE into the curriculum for first-year students (Echegoyen *et al.*, 2019; McCabe and Olimpo, 2020; Leyser-Whalen and Monteblanco, 2022). The University of Maryland, Baltimore County BUILD program developed a Course-based Undergraduate Research Experience (CURE) in addition to an online badge program (Ott *et al.*, 2020). The BUILD program at the University of Detroit Mercy developed a CURE metagenomics course in the winter of 2019, where students investigated the bacterial community composition at a local lake and reported gaining authentic research experience (Baker *et al.*, 2021). The national evaluation of the BUILD programs included tracking students who participated in novel curricula developed or revamped by BUILD. Over 10,000 students have enrolled in these courses across the BUILD sites.

Reforming outdated and ineffective curricula in STEMM disciplines has been an ongoing strategy in national efforts to broaden diversity in STEM[M] (AAU, 2013; AAC&U, 2023; Mack *et al.*, 2019; Talanquer, 2014). For example, the AAC&U established Project Kaleidoscope in 1989 and Teaching to Increase Diversity and Equity in STEM (TIDES) in 2014 to transform STEM[M] teaching and learning in higher education. Discipline-based education research (DBER) (National Research Council, 2012) has also expanded over the past two decades. Such efforts to address STEMM disparities via curriculum matter because they are revolutionizing not only what is being taught but how it is being taught by looking at efforts to engage an increasingly diverse student population. However, little is also known about whether these efforts, despite aims to improve learning for every student, are leading to similar outcomes for all groups of students. Thus, critical examination of a novel curriculum is warranted. The BUILD evaluation has been described as an equity-focused impact evaluation (Guerrero *et al.*, 2022; Maccalla *et al.*, 2022) because of its focus on determining programmatic effects on clearly defined outcomes (e.g., RSE) for diverse and often underrepresented groups (i.e., women, specific racial/ethnic groups, persons with disabilities) in biomedical fields (Maccalla *et al.*, 2023).[1] Employing Intersectionality as an analytic lens, we examine the differential impact of two program features: the BUILD Scholars and BUILD novel curricula. A systematic review of the literature on WOC in STEMM noted the need for national quantitative longitudinal studies (Ong *et al.*, 2011). Thus, we focus on WOC in STEMM to demonstrate how a QuantCrit approach can center on certain marginalized groups to gain a deeper understanding that can inform how to transform programs and conditions that contribute to disparities.

Critically Examining RSE

Derived from psychology, self-efficacy is concerned with one's judgment of one's competence in a domain (Bandura, 1977). RSE is conceptualized as the belief in one's ability to conduct research, an important skill for individuals who aspire to pursue STEMM-related careers. Self-efficacy is an increasingly important part of career readiness assessment (Betz and Borgen, 2000), including efforts to understand career development in science-related fields (Diversity Program Consortium, n.d.; Forester *et al.*, 2004). Evidence suggests that interventions that aim to enhance students' identity as scientists and their RSE have the potential to support retention on a science career pathway (Mullikin *et al.*, 2007; Maton *et al.*, 2016) and predict aspirations for a research career (Adedokun *et al.*, 2013). Disparities in self-efficacy, such as women reporting lower RSE than men, are also documented (Gibbs *et al.*, 2015). Lastly, RSE is a useful measure of understanding the extent to which science training interventions sustain individuals' interest in science-related research careers, particularly those from underrepresented groups (Bakken *et al.*, 2010). One study found that brief educational interventions can increase biomedical RSE for WOC compared with White women students (Bakken, *et al.*, 2010). However, less is known about program activities sustained over weeks or months or when program activities focus on improving the science curriculum. With an improved science course curriculum, the implicit hypothesis is that increased active learning strategies and research-related activities embedded in revamped or newly developed courses will increase RSE.

The four key sources of self-efficacy were originally theorized as mastery experience to develop skills, vicarious experiences (comparing one's experiences to others), verbal persuasion (encouragement or discouragement from others), and physiological feedback (self-perceptions of one's physical and

[1]For a complete understanding of how equity-based evaluation frameworks shaped the Enhance Diversity Study, please see p. 14 of (Guerrero *et al.*, 2022).

emotional cues) (Bandura, 1977). Additionally, Usher and Pajares (2008) suggest that while mastery experience is a prominent source of self-efficacy, the strength and influence of the sources differ depending on contextual factors such as gender, race/ethnicity, academic ability, and domain. Considering Intersectionality, examining students' perceptions of self-efficacy can help connect and unpack how socialization in a biomedical initiative may have a differential impact on groups based on their social position in hierarchies within society. In other words, self-perceptions of WOC are shaped by structural conditions. In a qualitative study of women in science and mathematics-related careers, Zeldin and Pajares (2000) found that vicarious experiences and verbal persuasions were instrumental sources for the development and maintenance of self-efficacy beliefs and that these beliefs are nourished by the relationships in women's lives and by the confidence that significant others express in their abilities. The study was limited to mostly White women already working in their STEMM careers. The present study is distinct in that it aims to measure the extent of differences in self-perceptions (i.e., RSE) for WOC undergraduates and interprets statistical differences as an indicator of how structural barriers shape differential patterns for WOC rather than differences attributed to individuals' cultural capital.

The CEC led efforts identified key outcomes of interest at critical training and career transition points; these outcomes are referred to as Hallmarks of Success (McCreath *et al.*, 2017) and include the development of RSE for undergraduate students (Cobian, 2019). Prior studies of BUILD sites demonstrated a strong positive effect of BUILD on RSE for first-year students (Cobian *et al.*, 2021; Crespi and Cobian, 2022) but did not examine outcomes for individuals' social locations with respect to race/ethnicity and gender identity. Syed *et al.* (2019) employed path analysis and found that self-efficacy affected science identity, with both affecting commitment to a science career, with no major differences by gender or race/ethnicity. Using data from the BUILD sites, we quantitatively examined RSE for WOC biomedical undergraduates, with particular interest in any differences for WOC who participated in BUILD activities such as the intensive BUILD Scholars program or BUILD-developed novel STEMM curriculum.

MATERIALS AND METHODS

Data Sources and Sample

We utilized a longitudinal survey dataset collected by the Enhance Diversity Study from the 10 BUILD programs at the 11 universities. The data are comprised of three survey instruments: the Higher Education Research Institute's (HERI) Freshman Survey (TFS), the Student Annual Follow-up Surveys (SAFS), and the HERI's College Senior Survey (CSS). The TFS collects baseline data on incoming first-time, first-year students. The TFS asks about student's precollege attitudes, perceptions, beliefs, and demographic data. The data includes four TFS administrations in the Fall of 2016, 2017, 2018, and 2019. The SAFS, developed by the CEC, is administered each spring starting in 2017. As a follow-up to the TFS, the SAFS collects students' perceptions and views on various educational and career goals and asks about college experiences. The dataset includes SAFS data from Spring 2017 to Spring 2022. Finally, graduating seniors complete the CSS (years 2018 through 2022 are included), which focuses on a broad range of college outcomes (e.g., academic achievement, satisfaction with college experience) and postcollege goals and plans. We merged the survey data with the BUILD program participation data to help us distinguish the students who participated in BUILD from those who did not participate in BUILD.

Our sample of 7250 students comprises four cohorts of first-year undergraduate students (2016–2019) who completed at least one follow-up survey (SAFS or CSS). The number of follow-up surveys for students ranges from one follow-up (32% of students) to five follow-up surveys (only 3% of students). Racial/ethnic and gender distributions for BUILD (1122) and non-BUILD students (6128) are depicted in Table 1. Among the 1122 BUILD students, 501 students participated only as Scholars and/or Associates (44%), and 919 (81%) students completed only the Novel Curriculum courses within BUILD sites. Demographic distributions for BUILD and non-BUILD students differ slightly, with 50% of BUILD students identifying as Black (17% for non-BUILD) and 21% as Latine (29% non-BUILD). Each sample is comprised of a larger proportion of women (76% BUILD and 66% non-BUILD) than men or nonbinary individuals. Among the BUILD students, 3% identified as American Indian and Alaska Native (AIAN) women, 12% identified as Asian American women, 40% identified as Black/African American women, 17% identified as Latine women, 0.45% identified as NHPI women, and 19% identified as White women.

TABLE 1. Sample sizes for student's race/ethnicity and gender by BUILD participation at baseline

	BUILD Students ($n = 1122$)		Non-BUILD Students ($n = 6128$)	
Variables	***n***	**%**	***n***	**%**
Race/Ethnicity[a]				
AIAN	41	3.65	155	2.53
Asian American	191	17.04	1500	24.51
Black/AA	556	49.59	1024	16.73
Latine	238	21.23	1749	29.00
NHPI	7	0.62	87	1.42
White	309	27.56	2460	40.19
Gender				
Women	853	76.02	4054	66.18
Men	255	22.64	2000	32.65
Non-Binary	15	1.34	71	1.16
AIAN Women	33	3.00	111	2.00
Asian American Women	138	12.30	939	15.32
Black/AA Women	450	40.11	735	12.00
Latine Women	189	16.84	1209	19.73
NHPI Women	5	0.45	61	1.00
White Women	212	18.89	1602	26.14

Note. BUILD students are those who participated as Scholars, Associates, and in Novel Curricula.

Abbreviations: AIAN = American Indian and/or Alaska Native; NHPI = Native Hawaiian and Pacific Islander.

[a]Race/ethnicity percentage distributions are inclusive of all races and/or ethnicities selected; therefore, total percentages exceed 100%, given respondents can select all that apply.

Variables

Dependent Variable: RSE. The dependent variable, RSE, is measured using a six-item scale that asks students about the extent to which they could complete the following tasks: 1) Use technical science skills, 2) Generate a research question, 3) Determine how to collect appropriate data, 4) Explain the results of a study, 5) Use scientific literature to guide research, and 6) Integrate results from multiple studies. Responses are on a 5-point Likert scale ranging from "not confident at all" to "absolutely confident." The item response theory (IRT) scaled score[2] of students' RSE is constructed using the graded-response model item parameters obtained from a national sample of the TFS 2016 survey from HERI. Using these item parameters, we constructed the Expected-A-Posterior (EAP) scores for the entire sample using the six items. For ease of interpretation of the score and to avoid negative RSE scores, we rescaled the score to have a mean of 50 and a SD of 10. RES scores range from 19.35 to 70.86 for the entire sample. We assessed the psychometric properties of this scale using the six items and conducted a differential item functioning analysis to ensure the scale functions similarly for all subgroups. This set of analyses allows us to conclude that the differences with respect to the outcome in the regression models when studying the differential impacts of BUILD on WOC are likely to be the true differences rather than a consequence of any measurement bias in the outcome.

Treatment Variable (BUILD Participation). We operationalized participation in BUILD by using the broadest definition of program engagement as our key treatment variable, which included *Scholars*, *Associates*, and those who participated in courses considered *Novel Curriculum*. BUILD *Scholars* are the most "intensely" treated and supported group of students, while *Associates* generally participate in a subset of BUILD activities (see Maccalla *et al.*, 2022, p. 61). *Novel Curriculum* is the most widely available aspect of the BUILD program (i.e., open-access undergraduate courses) and includes the largest proportion of undergraduate students at each BUILD site.

For our current analyses, we defined BUILD participation in two ways. First, we created an indicator variable for BUILD Scholars, which took a value of "1" for those who participated as Scholar and/or Associate and a value of "0" otherwise. Second, we created an indicator variable for BUILD Novel, which took a value of "1" for those who completed a course identified as Novel Curriculum and a value of "0" otherwise.

Cohorts of students have been admitted each academic year since the start of BUILD in 2015. Most students enter BUILD between June and August of each academic year to participate in the TFS, which closes in October. For the longitudinal analysis, we created a time-varying *BUILD Now* variable that indicated whether or not a student had been admitted to the BUILD program as a Scholar or Associate as of the time that they took a given survey; that is, *BUILD Now* takes a value of "1" at the time a student is admitted to the program and remains "1" thereafter. BUILD Now remains "0" for students never admitted to the BUILD program. This is useful because students enter the BUILD programs at different times of the year, and a time-varying variable helps us correctly capture the students affiliated with BUILD at a particular time. Similarly, we created a *Novel Now* variable, which allows us to correctly capture the students who took the BUILD-developed courses at times prior to their final reported RSE score.

Additional Explanatory Variables. We included additional variables at baseline in our propensity score model, including Pell Grant awardee status, first-generation college student status (neither parent nor guardian graduated college), high school grades, and incoming RSE. We also controlled for the ten BUILD programs. We recognize that these categories shape material outcomes for individuals: thus, we acknowledge and account for the evidence-based knowledge about barriers that low-income and first-generation college students encounter in postsecondary education.

Gender and Race/Ethnicity. Students provided self-reported information about their gender identity and racial and ethnic identity. We recognize that racial and ethnic categories do not simply describe or categorize women but aim to model how women college students' experiences are shaped by multiple systems of oppression (i.e., racism and sexism) in STEMM disciplines (Bowleg, 2008; Wong-Campbell and Ramrakhiani 2024).

We created six groups, as listed in Table 1. Women include cisgender and transwomen. To code race and ethnicity, we took a unique approach that aimed to account for students' multiracial identities, instead creating indicator variables with "1" indicating their self-reported racial and ethnic identity and "0" indicating they are not part of that race/ethnic group. For example, a multiracial Black and Asian American cisgender woman was coded as "1" for the Women variable and coded as "1" for Black and "1" for Asian American. Thus, our regression models do not force students into one discrete category for race/ethnicity. This is important for accounting for students' multiple self-reported racial/ethnic identities rather than the evaluator potentially erasing or masking students' voices as represented by their selections. Intersectionality acknowledges that power structures can operate on a group even when not numerically underrepresented. For example, while some STEMM equity efforts focus on groups deemed numerically underrepresented in STEMM (National Institutes of Health [NIH], 2019), we chose to include Asian American WOC, whose experiences are often not discussed in STEMM diversity efforts (Cobian *et al.*, 2022) yet still are negatively impacted by gendered racism.

Research Experience. The research experience variable comprises student responses to items from both CSS ("*Since entering college, have you participated in an undergraduate research program?*") and SAFS ("*In the past 12 months, have you had any opportunity to conduct your own scientific research or to participate in scientific research directed by others?*") surveys. Using these items, we created an indicator variable for research experience, which takes on a value of "1" for all students who responded "Yes" to either of these items and "0" if the response is "No" to both. Another CSS item asks students, "*How many months since entering college (including summer), did you work on a professor's research project?*" (response options include

[2]Please refer to Thissen and Orlando (2001) for more details.

0 months, 1–3 months, 4–6 months, 7–12 months, 13–24 months, and 25+ months). Students who indicated a nonzero number of months to this item were also coded as "1."

Analyses

To examine the differences in RSE scores across gender and race/ethnicity groups for BUILD-related activities, we conducted a series of analyses. First, we analyzed the data using descriptive statistics (RSE mean and SD) at college entry (i.e., baseline) and the percentage of students by subgroups in the sample. Next, we examined whether the BUILD Scholars and broader BUILD Novel Curriculum activities worked differently for WOC groups.

Propensity Score Matching. We used a propensity score approach to account for the selection bias between the BUILD students and those who did not participate in the BUILD programs (non-BUILD students). Note, for the propensity score matching, we employed a broad definition of BUILD, which included all biomedical undergraduates who were either BUILD Scholars and/or participated in the BUILD-developed novel curricula. The propensity score is a probabilistic score that is a scalar summary comprising all the key observed covariates that one could match on (Rosenbaum and Rubin, 1985). These scores allow us to create groups of students in treatment (BUILD) and control (non-BUILD) conditions who are similar on all key covariates so that the sample distributions are identical across groups (Rosenbaum and Rubin, 1985), thus reducing the confounding effects (Stuart, 2010). Lastly, less than 5% of data were missing across gender and racial/ethnic groups. All missing cases were listwise deleted, and complete cases were retained for propensity score matching and the final outcome models.

First, we conducted a propensity score estimation using the MatchIt package in *R* (R Core team, 2022). We used the full matching approach where every student in the treatment group is matched to at least one student in the control group, and every student in the control group is matched to at least one student in treatment (Hansen, 2004; Stuart and Green, 2008). For the propensity score estimation model, the outcome variable was a binary indicator variable of whether the student had ever participated in the BUILD program. The additional covariates were measured prior to treatment (see Supplemental Table SA1 in the appendix for a list of covariates). We included the RSE score at baseline as a pretest variable to account for students' incoming RSE. We estimated the ATE, which is the potential effect of the treatment on the entire population (Guo and Fraser, 2014).

After matching on key covariates, we examined the covariate balance by assessing the density plots, balance tables, and the quality of the match by checking whether any students were unmatched. The balance plot (see Figure 2) summarizes the balance across the covariates using the Absolute Standardized Mean Difference (SMD). The SMD is the difference in the means of each covariate between treatment groups standardized by a factor that is on the same scale for all covariates. The standardization factor is the SD of the covariate in the pooled SD across both groups when using the ATE estimand. An absolute SMD close to 0 indicates a good balance. Recommended absolute values are those less than 0.1 (Greifer, 2022).

Outcome Models. Second, we fit a series of multiple regression models with the outcome in the dataset to examine the treatment (i.e., intervention) effects of BUILD activities and their impact on RSE for WOC. A key feature of our QuantCrit analytic approach is that we fit a different outcome model for every racial and ethnic group of women. Considering the longitudinal nature of the surveys, the outcome models use each participant's RSE score 3 years (after college entry) after the intervention. We narrowed our analyses to one timepoint post-intervention for two reasons: First, our goal is to illustrate a QuantCrit approach in an evaluation setting by focusing on intersectionality and the effects of BUILD-related activities on students who have historically been excluded from full participation in STEMM. Second, we were not focused on the growth or change of the student's RSE over time, but instead were interested in understanding student's RSE when they are involved in BUILD-related activities and have spent significant time at their institution while including baseline characteristics at college entry to control for any selection bias.

To examine the differential impact of BUILD activities on WOC, we fit stepwise multiple regression models for every racial/ethnic group (e.g., Asian American, Latine, Black/AA) post-intervention using the total sample of undergraduates. In other words, we ran regression models one at a time for each of the racial/ethnic groups as the key variable of interest to obtain the coefficients for that particular group. This allowed us to keep the statistical power of the total sample of undergraduates while still understanding how coefficients changed for each racial/ethnic group and the interaction terms. This modeling differs from other regression models where there is a reference group for the race/ethnicity variable. We included interactions if we saw a significant main effect for a particular race/ethnicity group. Variables were included one by one with ATE weighting after the propensity score estimation. We included time-varying indicators for BUILD participation—BUILD Now and Novel Now, as well as the RSE scores at baseline to account for students' RSE at college entry. Next, we added gender, each ethnoracial group, and the Research Experience variable to account for the influence of URE participation on RSE. We added the specific ethnoracial group by women interaction to model multiple and simultaneous social locations of race and gender for the WOC subgroups. Lastly, we added the interactions of the BUILD x WOC to examine whether BUILD differentially impacted WOC and included the fixed effects for sites to control for any differences across the BUILD sites. Due to the large number of models that were fit in a stepwise fashion across the various racial/ethnic groups, in Tables 2 and 3, we only present the final model in the stepwise regression for each racial/ethnic group.

Positionality

We believe critical quantitative research requires self-reflexivity and transparency (Rios-Aguilar, 2014). An active and authentic positionality statement acknowledges the centrality of racism in society and allows researchers to explore and navigate their own connections and biases related to the work through self-reflexivity (Milner, 2007; Suzuki *et al.*, 2021; Castillo and Gillborn, 2023). The first author is a South Asian WOC who earned her bachelor's and master's degrees in physics. Growing up in a society that focused heavily on

TABLE 2. ATE results for the gender, race/ethnicity, and intersectionality subgroups with RSE as the outcome variable and the conditional effects of BUILD scholars for WOC

Variables	Model for AIAN (A) Estimate (SE)	Model for Asian American (B) Estimate (SE)	Model for Black/AA (C) Estimate (SE)	Model for Latine (D) Estimate (SE)	Model for NHPI (E) Estimate (SE)	Model for White (F) Estimate (SE)
Intercept	25.26 (0.92)***	25.48 (0.96)***	25.7 (0.92)***	25.25 (0.94)***	25.41 (0.92)***	25.00 (0.93)***
RSE at Baseline	0.41 (0.02)***	0.41 (0.02)***	0.41 (0.02)***	0.41 (0.02)***	0.40 (0.02)***	0.40 (0.02)***
BUILD Now[a]	1.69 (0.58)**	1.52 (0.62)*	1.30 (0.59)*	2.12 (0.60)***	1.35 (0.58)*	2.15 (0.64)***
Novel Now[b]	1.98 (0.60)***	1.97 (0.60)**	1.84 (0.60)**	1.91 (0.60)**	2.01 (0.60)***	2.06 (0.60)***
Research Experience	2.82 (0.36)***	2.82 (0.36)***	2.81 (0.36)***	2.83 (0.36)***	2.78 (0.36)***	2.85 (0.36)***
Ethnoracial Group[c]	5.62 (1.81)**	−0.04 (0.61)	−0.73 (0.76)	0.52 (0.65)	6.03 (1.80)***	0.94 (0.55)
Women	0.08 (0.34)	−0.08 (0.39)	−0.37 (0.36)	0.12 (0.39)	0.08 (0.34)	0.58 (0.43)
Non-Binary	−0.78 (1.81)	−0.93 (1.81)	−0.72 (1.81)	−0.85 (1.81)	−0.82 (1.80)	−0.77 (1.81)
Ethnoracial Group-Women Interaction	−5.26 (2.04)*	0.05 (0.76)	1.37 (0.89)	−0.21 (0.75)	−8.1 (2.49)**	−1.12 (0.69)
Ethnoracial Group-Women-BUILD Now Interaction	−4.38 (2.29)	0.46 (1.24)	2.95 (1.35)*	−3.88 (1.27)**	5.63 (2.92)	−2.01 (0.95)*

Note. Each model includes the total sample of students. Each model includes site as a fixed effect.
[a]BUILD Now refers to Scholar/Associates, a time-varying measure that captures the average effect of BUILD exposure up to the first three timepoints.
[b]Novel Now refers to students' exposure to novel curriculum, a time-varying measure.
[c]Ethnoracial Group refers to each of the racial/ethnic subgroup as noted at the top of the columns. For example, the estimate of −5.26 in the first column corresponds to those who identify as AIAN, and an estimate of −4.38 (last row) corresponds to AIAN women-BUILD interaction.
*$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$.

the caste of a person and less on educational opportunities for all influenced her research interests to study equity disparities for historically disadvantaged groups. The second author is a multiracial cisgender WOC. Early experiences of marginalization in math and science throughout her educational trajectory influenced research interests in identifying systemic barriers that multiply-marginalized individuals experience in STEMM education. The third author is a multiracial, nonbinary, first-generation college student/professional whose experience with adversity has influenced their social and scholarly commitments to equity. The fourth author is a queer-identified scholar of evaluation deeply committed to engaging in work intended to promote social justice and change. As critical researchers, we know well that our experiences and identities shaped our decisions to center the analyses presented in this study on women and underrepresented groups, with the goal of centering and revealing differential experiences of BUILD students and to develop implications and future directions for research that counters decades of STEMM mainstream research that has often reproduced class, race, and gender oppression (Kincheloe and McLaren, 1994; Stage, 2007).

RESULTS

To answer the first research question, we examined students' demographic characteristics in the sample (see Table 1). Students were asked about RSE in the fall of their first year of

TABLE 3. ATE results for the gender, race/ethnicity, and intersectionality subgroups with RSE as the outcome variable and the conditional effects of novel curriculum exposure for WOC.

Variables	Model for AIAN (A) Estimate (SE)	Model for Asian American (B) Estimate (SE)	Model for Black/AA (C) Estimate (SE)	Model for Latine (D) Estimate (SE)	Model for NHPI (E) Estimate (SE)	Model for White (F) Estimate (SE)
Intercept	25.29 (0.92)***	25.43 (0.96)***	25.7 (0.92)***	25.24 (0.94)***	25.4 (0.91)***	25.15 (0.93)***
RSE at Baseline	0.40 (0.02)***	0.40 (0.02)***	0.40 (0.02)***	0.40 (0.02)***	0.40 (0.02)***	0.40 (0.02)***
BUILD Now[a]	1.28 (0.58)*	1.35 (0.58)*	1.35 (0.58)*	1.26 (0.58)*	1.20 (0.58)*	1.33 (0.58)*
Novel Now[b]	1.84 (0.61)**	1.97 (0.63)**	1.60 (0.65)*	2.29 (0.64)***	1.68 (0.60)**	1.96 (0.66)**
Research Experience	3.00 (0.36)***	2.98 (0.36)***	2.98 (0.36)***	2.98 (0.36)***	3.03 (0.36)***	3.03 (0.36)***
Ethnoracial Group[c]	4.14 (1.63)*	0.09 (0.61)	-0.52 (0.78)	0.41 (0.64)	6.47 (1.83)***	0.63 (0.55)
Women	0.21 (0.34)	0.10 (0.39)	-0.22 (0.36)	0.26 (0.39)	0.23 (0.34)	0.63 (0.43)
Non-Binary	−0.51 (1.71)	−0.64 (1.71)	−0.41 (1.71)	−0.68 (1.71)	−0.51 (1.7)	−0.64 (1.71)
Ethnoracial Group-Women Interaction	−4.59 (2.10)*	0.03 (0.76)	1.59 (0.91)	−0.32 (0.75)	−9.49 (2.42)***	−1.29 (0.68)
Ethnoracial Group-Women-Novel Now Interaction	0.84 (2.95)	−0.93 (1.43)	1.15 (1.21)	−2.72 (1.27)*	16.26 (3.64)***	−0.34 (1.00)

Note. Each model includes the total sample of students. Each model includes site as a fixed effect.
[a]BUILD Now refers to Scholar/Associates, a time-varying measure that captures the average effect of BUILD exposure up to the first three timepoints.
[b]Novel Now refers to students' exposure to novel curriculum, a time-varying measure.
[c]Ethnoracial Group refers to each of the racial/ethnic subgroup as noted at the top of the columns.
*$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$.

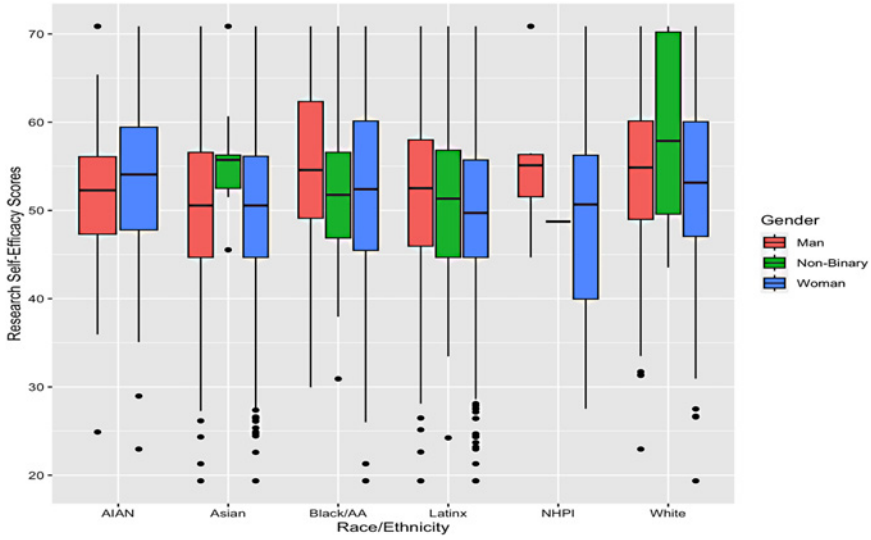

FIGURE 1. RSE by gender and race/ethnicity for all students at baseline. *Note*. Because RSE is an IRT-scaled score, the minimum value across the various groups is 19.35, and the maximum value is 70.86, except as follows: for the nonbinary gender category, the minimum value is 24.23, for the AIAN subgroup, the minimum value is 22.95, and for the NHPI subgroup, the minimum value is 27.52.

college, thus providing a baseline RSE score before college experiences and/or participation in BUILD-related activities. For all incoming first-year students in the sample, the mean RSE score at baseline was 52.52, with a SD of 9.97. We also examined baseline RSE for intersectional groups to examine differences by combined gender and race/ethnicity (Figure 1). At college entry, mean RSE scores were statistically significantly lower ($p < .05$) for both Latine and Native Hawaiian and Pacific Islander (NHPI) women compared with their men counterparts. In contrast, AIAN and Asian women had a statistically significantly higher mean RSE score than their men counterparts. We found no statistically significant differences in the RSE scores for Black/AA women and White women.

To assess the quality of the propensity score approach, Figure 2 shows that the absolute SMD is less than 0.1 for most covariates, indicating a good balance among BUILD and non-BUILD students. This allows us to draw valid inferences from our outcome models regarding BUILD students. The balance table is presented in the appendix (Supplemental Table SA1) which includes the means of the covariates in the treated group (BUILD students), control group (non-BUILD students), and the SMDs in the sample before and after matching. Figure 3 visualizes the propensity scores distribution of those matched using a jitter plot. The plot indicates that there were no unmatched students. Density plots among the various sub-groups for the BUILD and non-BUILD students were examined to confirm no imbalances (see Supplemental Appendix, Figure SA1).

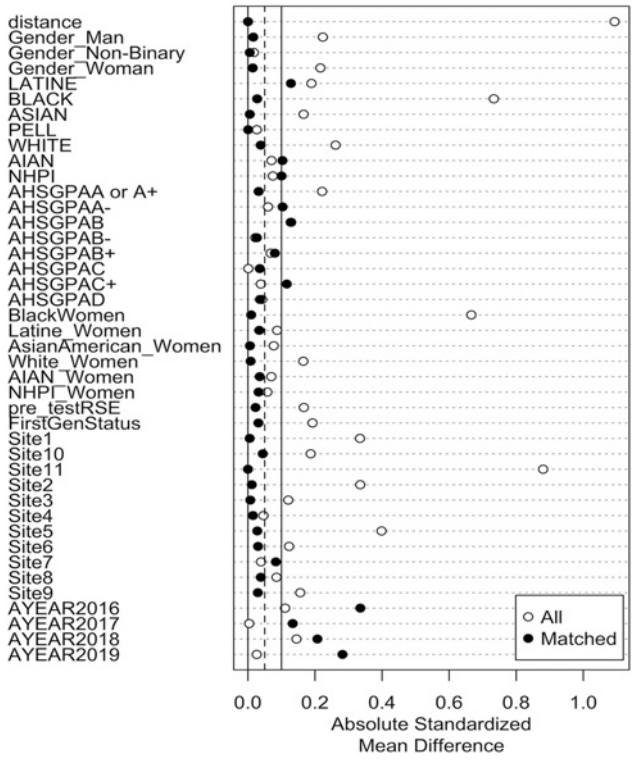

FIGURE 2. Balance of covariates before and after weighting.

Predicting RSE without Intersectionality

We found that the incoming RSE scores have strong predictive power across all models (estimate of 0.41, $p < 0.001$). This can be interpreted as students with higher RSE at the start of their first year tend to have higher RSE later in their undergraduate studies. Prior to accounting for gender, research experience, and sites, results indicate that participation as a BUILD Scholar significantly contributes to higher RSE. Additionally, participating in undergraduate research strongly predicted RSE, attenuating the effect of being a BUILD Scholar. In other words, undergraduates who reported participation in URE, inside or outside BUILD programs, reported higher self-efficacy in their research skills. Left at this stage, evaluators might conclude that the program works to increase RSE for all undergraduates.

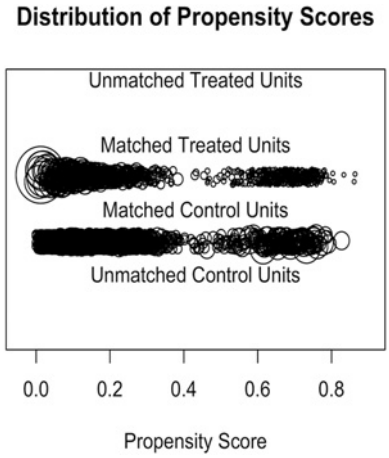

FIGURE 3. Quality of matching depicted by the distribution of the propensity scores.

Approaching Intersectionality: Differential Effects of BUILD Activities for WOC

To answer our second research question, Table 2 present Models A to F for each race/ethnicity (or ethnoracial) group denoted at the top of the columns. The model for each race/ethnicity group compares that subgroup with all other students. Students who had become BUILD Scholars and students who participated in the novel curriculum had statistically significantly higher RSE scores compared with their non-BUILD counterparts after controlling for RSE at college entry, research experience, and BUILD site. For example, those BUILD Scholars who identified as Black/AA (Model C) had 1.3 points increase in RSE and those who identified as White (Model F) had 2.15 points increase in RSE compared with their non-BUILD counterparts. After identifying when it was appropriate to examine interaction effects for each model, we found that BUILD Scholars who identified as Black Women (Model C) and White women (Model F) were more likely to have higher RSE scores, whereas Latine Women (Model D) were more likely to have lower RSE scores (each group is compared with every other non-BUILD undergraduate in the sample) (see Table 2). To understand the conditional effects of BUILD on RSE scores, we used the estimate of the ethnoracial Group-Women-BUILD Now interaction term across the specific racial/ethnic subgroups (last row of Table 2). Equation 1 helps us understand the estimates for the variables in our model for Black/AA group (Model C from Table 2); holding constant all other variables, Black/AA Women who are BUILD Scholars have a 4.25 ($1.30 + 2.95 = 4.25$) increase in their RSE score. Similarly, in Model D, Latine Women who participated as BUILD Scholars show a decrease of 1.76 points in their RSE score ($2.12 - 3.88 = -1.76$), and White women who participated as BUILD Scholars have a 0.14 increase in their RSE score.

$$\begin{aligned} \text{RSE}_{\text{scores}} = {} & 25.7 + 1.30\,\textit{BUILDNow} + 1.84\,\textit{NovelNow} \\ & + 0.41\,\textit{RSE Pretest} + 2.81\,\textit{Research Experience} \\ & + (-0.37)\,\textit{Women} + (-0.73)\,\textit{Black/AA} \\ & + 1.37\,\textit{Black/AA Women} \\ & + 2.95\,\textit{BUILDNow*Black/AA Women} \end{aligned} \quad (1)$$

Furthermore, AIAN women (Model A) and NHPI women (Model E) who were BUILD Scholars had lower RSE scores by 5.26 points and 8.1 points, respectively, in comparison to their non-BUILD counterparts (Table 2). The small sample sizes for NHPI students means that researchers must interpret the estimates with caution. We felt it was important to still include these students in the analyses rather than remove NHPI students from the regression model (thereby erasing their responses and experiences from the analysis). A QuantCrit approach can encourage more research in this area and advance methods and analysis for groups with smaller sample sizes. We return to this in the *Discussion* section.

To answer our third research question, whether there is a differential effect of participation in STEM novel curricular courses on RSE for WOC, we present results for each race/ethnicity group (Table 3, models A to F). Students who participated in BUILD-developed curricula had statistically significantly higher RSE scores compared with their non-BUILD counterparts. Participation in undergraduate research and RSE at college entry (pretest scores) were both strong predictors of RSE, attenuating the effects of participating in BUILD-developed curricula. We found that Latine Women have a lower RSE compared with other biomedical undergraduate students. However, the overall conditional effects of Novel Curriculum are less pronounced than the conditional effects of participation as a BUILD Scholar across all WOC groups. The lower intensity of the engagement in a novel STEMM curriculum course compared with a scholar program may explain this, but we provide additional context in the Discussion section.

Further Examination of Black and Latine Women

Based on the results above, we decided to understand and contextualize the statistically significant results for Black and Latine Women involved in the BUILD initiative. Hence, we conducted further analyses that accounted for context. Most BUILD sites are teaching-focused institutions, and several sites are also Minority-Serving Institutions (MSIs). With this context in mind and our positionalities as higher education scholars and evaluators, we considered whether the complexity of gendered racism at MSIs may be contributing to the higher RSE for Black women in BUILD and lower RSE for Latine women in BUILD. The inclusion of BUILD sites as fixed effects in the regression models indicated that some sites were significantly different from each other with respect to students' RSE. Reviewing descriptives of the sample, we confirmed that most Black women undergraduates were enrolled at the two Historically Black Colleges and Universities (HBCUs) in the study, and Latine women were primarily enrolled at four institutions, all with Hispanic-serving institution (HSI) or joint HSI/AANAPISI (Asian American and Native American Pacific Islander) designations.

We ran ANOVA tests to determine differences in RSE for Black women and/or Latine women across the BUILD sites. Examining Black women, we found significant differences between two BUILD sites. The RSE for Black women at a private HBCU was 6 points ($p < 0.05$) higher than a public HSI site (Note: The SD of the RSE scale is 10). Similarly, we found that one of the HSIs with a large population of Latine students (over 80%) had Latine women BUILD Scholars with RSE

scores 4 points higher than their counterparts at a West Coast HSI BUILD site. We discuss the implications of differential outcomes in the *Discussion*.

Limitations

Our study has a few limitations, which have been noted as common in other studies using the QuantCrit approach. First, inferences regarding the AIAN and NHPI groups are limited by the small sample sizes for these groups (Castillo and Gillborn (2023) and Suzuki *et al.* (2021) note similar challenges). These results (Models A and E in Tables 2 and 3) must be interpreted with caution because of the small sample sizes. Still, we chose to prioritize the inclusion of every undergraduate in the sample rather than erasing students who identified as AIAN and NHPI. To this end, it is important to consider ways to include these groups in all analyses or oversample subgroups, like for Native Hawaiian and Pacific Islander students (Teranishi *et al.*, 2013) and Indigenous students (Lopez and Gaskin, 2022).

From a methodological standpoint, Bayesian methods that are sensitive to small sample sizes and not driven by large-sample theory can be useful in these cases. Given that there were only five BUILD NHPI women in the sample, we examined the raw data and descriptive statistics for these students and found them to have a high RSE score with a mean of 56.97 in comparison to 47.76 for those who are non-BUILD NHPI women. We are mindful of the implications that we make of these numbers due to our positionality as outsiders from the NHPI and Indigenous experience, as well as the small sample sizes. Examining any patterns or trends from these students is at least one step that can provide additional insight into what might be contributing to differences.

Associated with the QuantCrit principle of centralizing race and racism, we point to the structural barriers of racism, colonialism, and settler colonialism that have contributed to the historical exclusion of Indigenous and Pacific Islander people from participation in postsecondary education. Rather than contributing to their erasure, a QuantCrit approach would invest in additional time and resources (as a next step) to understand the impact of a program for these groups. For example, additional qualitative data analysis, mixed methods approach, and critical reflective conversations with the programs and participants can help contextualize and/or provide insight into our inferences.

Another limitation is the broad categorization of race/ethnicity (a challenge noted by Castillo and Gillborn (2023) and Suzuki *et al.* (2021) in previous studies). The survey asked students which race/ethnicity category described them the best and provided options that aggregated and conflated race and ethnicity. In some surveys in the later years, students were given options to select subcategories for a specific race group; for example, if they marked Asian, students could further select Chinese, Filipino, Korean, or Indian as specific ethnic identities. However, these subitems had extremely small sample sizes, and the lack of these items in survey years 2016 to 2019 meant that we had to use the broad race categories. In program evaluation, if we want to know "for whom" the intervention works and if there are differential effects of a program for specific sub-groups, researchers need to be mindful of how race and ethnicity are collected and how other social identities are asked to survey respondents and how they inform a particular axis of power. Lastly, the current Enhance Diversity surveys do not have direct measures of racism or gendered racism as this was not a primary focus of the broader evaluation. This study is an example of how to approach QuantCrit even when data sources may be limited in addressing race and racism. As a result, we were careful in our interpretation and implications of the regression models and their capacity to directly measure structural inequality. In our discussion section, we present how, using existing datasets, we infuse and embed critical reflection throughout our work, from framing research questions to analyses to interpretation of the findings.

DISCUSSIONS AND IMPLICATION

This study aimed to utilize QuantCrit and Intersectionality to better understand outcomes from a national biomedical initiative. We acknowledge that racism is pervasive among us and thus incorporated an anti-racist orientation while we developed our research questions, conducted statistical analysis, and interpreted results. We contextualize and provide a QuantCrit interpretation of results and implications for next steps in program evaluation that can further advance QuantCrit.

Contextualizing Results

This study considered two main BUILD-related activities—participation as a BUILD Scholars and completion of a BUILD-developed Novel Curriculum course in STEMM disciplines. After using propensity score matching to create similar groups of students for whom we could draw sound inferences, we found that students who participated in both BUILD-related activities (both in Scholars/Associates program and Novel Curricula) reported significantly stronger RSE scores relative to their non-BUILD counterparts. A typical evaluation might conclude that the programs work, but not delve deeper into potential disparities in outcomes and working with program facilitators to enact social change that addresses those disparities.

Employing a QuantCrit approach, we found differential effects of the BUILD Scholars program for Black, Latine, and White women. We also found a differential effect for students who completed a Novel Curriculum course (BUILD Novel) and identified as Latine women and/or NHPI women. With respect to increased RSE for Black women, these results are encouraging. They suggest that STEMM training programs that provide a combination of financial support, URE, and mentorship can contribute to increased self-perceptions of one's ability to engage in scientific research for Black women in STEMM. However, there were no statistically significant differences for the larger group of Black women undergraduates enrolled in Novel STEMM Curriculum. This suggests that Black women benefit the same as others from new STEMM curricula. Further, Latine women were more likely to have lower RSE after participating in new STEMM curricula. Thus, if national goals are to enhance participation from groups who are underrepresented in STEMM, then this study suggests that the curricular changes do not significantly enhance

RSE for WOC underrepresented in STEMM. Further, Latine women who participated in the BUILD Scholars program had decreased RSE compared with all other non-BUILD biomedical students. Overall, these findings suggest that marginalization looks different for women depending on their social locations with respect to racial/ethnic identity, thus underscoring the importance of considering Intersectionality and cautioning against approaches that analyze WOC as a single group (essentialism).

Considering QuantCrit while also examining a psychological construct like RSE, one's perception of their efficacy in research is partly shaped by structural inequality, not individual (in)adequacy (Usher and Pajares, 2008). The findings substantiate prior literature suggesting that within and across groups, individuals may weigh the four key sources of self-efficacy differently (Bandura, 1997; Usher and Pajares, 2008). Because the implementation of the BUILD program varied, we cannot determine the extent to which each site focused on attending to key sources of self-efficacy. While short-term self-efficacy interventions can increase RSE for WOC (Bakken, *et al.*, 2010), these findings suggest that grant-funded STEMM intervention programs might need to place additional targeted emphasis on self-efficacy to support Latine women students and potentially other WOC groups.

One defining QuantCrit principle recognizes that numbers alone do not tell the whole story (Garcia *et al.*, 2018; Gillborn *et al.*, 2018, Suzuki *et al.*, 2021; Castillo and Gillborn, 2023). Results from quantitative analyses should be contextualized with researchers' critical sensibilities, and additional research related to the analysis should be presented when available. Considering the differing results for Black and Latine women in the intensive BUILD Scholars program, a QuantCrit approach requires interrogating the structures and conditions that shaped these differences in student's RSE. BUILD-awarded sites included HBCUs and HSIs. Thus, we must consider how these campuses may have already been positioned to support the success of STEMM-engaged WOC. For example, HBCUs have a well-documented legacy of developing the talent of their students (Owens *et al.*, 2012). On the other hand, the unique history of the emergence of HSIs in the United States, and the heterogeneity among them has led scholars to examine their "servingness" (Garcia *et al.*, 2019). Connecting to the results of different RSE scores between the BUILD sites that were also HSIs, our findings support the idea that some HSIs do a better job of serving their students from underrepresented groups compared with others. Considering Intersectionality, future program evaluation (particularly at HSI sites) can take a critical approach to scholarship on HSIs, or "intersectional servingness" (Garcia and Cuellar, 2023) to examine other axes of structural inequality and the potential of HSIs to address the wide range of students whom they serve.

Additionally, BUILD sites developed various curricula, which raises the question of what types of courses and pedagogical innovations may or may not be effective for groups based on their social locations in relation to race, gender, class, etc. Future evaluation studies might incorporate qualitative data or integrate local evaluation findings to interrogate the findings and work in partnership with grant awardees to think critically about transformative change (Mertens, 1999).

Implications for Studying STEMM Diversity Efforts

This study provides an example of using a QuantCrit approach in a large-scale, federally funded program evaluation context. Using an Intersectionality lens will allow programs to create and refine interventions that work for groups often erased or decentered from large-scale evaluation studies. The interventions can be tailored to support specific groups (e.g., Black Women or Latine Women) entering and sustaining the biomedical workforce. This work addresses the groups for which we and the larger Diversity Program Consortium hope to see meaningful changes and can be extended to other important underrepresented groups (Maccalla *et al.*, 2023) in future studies.

The Importance of "Using Findings to Facilitate Social Transformation". The research/evaluation process encouraged us to think, "What would we do in a QuantCrit evaluation setting that might differ from a traditional research setting?" Suzuki and colleagues (2021) posit the three "moments" (development of research questions and variables, the role of race in analytic models, and interpretation of results through a critical framework) where researcher influence is essential in infusing principles of QuantCrit. We believe QuantCrit in program evaluation lends itself to a "Moment 4 – Using Findings to Facilitate Social Transformation," where evaluators can help facilitate discussions about the role of racism in explaining differential outcomes and identifying potential solutions. This moves toward the equity-oriented fifth tenet of CRT of using numbers to advance social justice (Gillborn *et al.*, 2018; Castillo and Gillborn, 2023). An anti-racist orientation can spur discussions in the local context about what structures, systems, policies, and practices may need to be changed. These critical-reflective discussions need not aim at the federal level but can impact the local communities or institutions. Evaluators can facilitate discussions by posing questions such as: How do we make sense of these findings within the power structures of our labs/departments/institution? What policies in our program(s) might contribute to these differential outcomes, and what can be done about them? What important needs and perspectives might the students in the program have to share? This calls for evaluators to embrace their role as partners in the evaluation process and agents of social change. These discussions can play a crucial role in guiding *new* research and evaluation questions, the role of race in the planned analyses, the interpretation of results, and subsequent programmatic decisions, moving toward a more equitable society with each inquiry cycle.

Implications for Study Design and Analytic Approach. A main contribution of this study is our analytic approach to accounting for race and ethnicity when employing regression analysis. QuantCrit evaluation can pay "careful" attention to the research questions and center our analyses around the simultaneous experiences of racism and sexism within STEMM. Race/ethnicity can be understood as an aggregate of a heterogeneous range of experiences and identities. Fitting a separate model for every race/ethnicity group allowed us to preserve the unique self-reported racial/ethnic identity of every student. Additionally, this approach reduced the number of times we compared one race group with another (typically

using White individuals as the reference group). Centering marginalized aspects of identity, rather than just controlling for them, may provide unique insight into a multitude of experiences within the same program. The answers we discover may not be as simple as single objective truths, such as "yes/no it worked," but rather a more complex understanding of multiple subjective realities operating in one time and space due to societal and institutional systems and structures at play.

Efforts to address underrepresentation in higher education or grant funding at later research career stages (Chen *et al.*, 2022) begin with ensuring interventions are having an impact within groups that are experiencing systemic barriers (e.g., WOC). Evaluation norms that value generalizability can take away from what may be gained from learning about specific populations (Teranishi, 2007; Gillborn *et al.*, 2018). Thoroughly examining evaluation data by subgroups can help identify and explain inequities and help researchers think more carefully about what will advance our understanding of the support various underrepresented groups need. Large-scale evaluations may not always have qualitative data for a particular group, but surveys and data analyses that use critical sensibilities can develop knowledge about how structural racism operates.

We recognize that this work can be complex. Large sample sizes and detailed information about programs and participants are needed to pursue disaggregated analyses while maintaining statistical power. Administering various surveys year after year across all the BUILD sites along with conducting cases studies provided us with a rich dataset, which we will continue using to answer questions about the effects of these programs. However, even with these resources, we faced challenges with small sample sizes for several racial/ethnic groups and therefore encouraged the interpretation of smaller subgroup findings with caution. When factoring in Intersectionality, sample sizes get even smaller. Additional research is needed for the AIAN and NHPI groups in this sample.

Researchers can plan to oversample individuals from various underrepresented groups during data collection and be mindful and intentional with the demographic data they plan to collect and incorporate recruitment campaigns and efforts during data collection that are sensitive to the historical mistrust of minoritized groups. Evaluation teams might also incorporate staff and researchers from the groups the program aims to target to learn from their expertise and cultural knowledge. This information can be helpful in the evaluation design, data collection, analyses, interpretation, and action stages of program evaluation. Additionally, other identities and systems of power, such as socioeconomic status or sexual identity, might be important to examine (Ghabrial and Ross, 2018). While we limited our area of foci to intersections of race/racism and gender/sexism to demonstrate a QuantCrit evaluation here, we recommend that evaluators and researchers examine other important domains of systemic marginalization.

Our study results point to the continued need for inquiry and analysis of STEMM scholar programs and STEMM classrooms to understand how and why outcomes might vary for different subgroups. This is especially important as STEMM departments nationwide continue efforts to revamp curricula and incorporate evidence-based curricular approaches to STEMM training, such as CUREs, POGIL, active learning practices, Learning Assistants, etc. It is important to consider how an intervention in a classroom (e.g., CUREs) can assess whether it produces outcomes that indicate further marginalization of a group (i.e., unintended negative consequences).

Whether employing QuantCrit or other critical approaches to quantitative research and evaluation, Biology education researchers conducting their own independent studies examining biology education with the aim of enhancing diversity in STEMM fields can employ the growing body of guidance on how to address systemic inequity in STEMM training (Pearson *et al.*, 2022). A QuantCrit approach can be useful for investigators conducting these studies, especially those examining outcomes with a critical lens that centers on social justice in its design, analysis, and interpretation of the study. Based on our work, we offer a few tips to help achieve this:

1. A QuantCrit approach to evaluating activities implemented through grant funding may reveal inequities (and positive outcomes) that otherwise might have gone unnoticed, unquestioned, or unchallenged by the researchers, funders, or other audiences. We suggest partnering with external evaluators and/or social scientists familiar with critical research paradigms, anti-racism efforts in STEMM, Ethnic Studies, and ideally QuantCrit research. An important step in this process is the critical reflection and discussions that will drive both the research/evaluation decisions and programmatic/institutional next steps.
2. QuantCrit does not necessarily need to involve complex statistical modeling. It can be used by incorporating a critical focus on understanding and questioning systems of power when examining quantitative data related to a program and outcomes.
3. For evaluation in particular, findings can be utilized as a tool to foster social transformation by facilitating discussions about the role of racism at campuses and within student experiences, and finally identifying potential solutions.

CONCLUSION

The BUILD program and corresponding national evaluation, with its diverse programmatic approaches, settings, and large sample sizes (Guerrero *et al.*, 2022), provides an opportunity to quantitatively disentangle what works, for whom, and under what contexts through a QuantCrit approach. Employing a critical approach to program evaluation that centers analyses of the differential effects of a program is particularly important for programs that aim to broaden participation in STEMM because they begin with acknowledging the large body of literature that emphasizes racism, sexism, and other systems of domination exist (McGee, 2020), rather than a neutral approach that overemphasizes individual STEMM trainees and underemphasizes the contexts, conditions, and social relations that shape individuals' outcomes.

We argue that complex problems require complex solutions, and critical analytic approaches allow for the complexity of experience and differential effects to be captured. When we understand more about participant experiences and contextualize quantitative findings with critical sensibilities that interrogate power structures, we can establish more

precision in our understanding of treatment effects of efforts aimed at dismantling the conditions that continue to reproduce inequity in STEMM (Pearson *et al.*, 2022). By moving beyond straightforward, traditional methods, quantitative researchers/evaluators can tell more complex and meaningful evidence-based stories. It is our responsibility to drill down on the multitude of experiences with a critical lens to ensure that evaluation and its resulting findings thoughtfully support the populations they aim to serve.

ACKNOWLEDGMENTS

The research team would like to thank members of the DPC for their steadfast contributions and all Enhance Diversity Study participants for their willingness to contribute. Furthermore, we would like to thank Ana Romero, Hector Ramos, Kenneth Gibbs, and Christa Reynolds from the DPC for their feedback. We would also like to thank the two anonymous reviewers for their feedback on the previous versions of this article.

REFERENCES

Adedokun, O. A., Bessenbacher, A. B., Parker, L. C., Kirkham, L. L., & Burgess, W. D. (2013). Research skills and STEM undergraduate research students' aspirations for research careers: Mediating effects of research self-efficacy. *Journal of Research in Science Teaching*, *50*(8), 940–951.

Alkin, M. C.& Vo, A. T. (2018). *Evaluation Essentials: From A to Z* (Second) New York: The Guilford press.

Association of American Colleges and Universities (AACU). (2023). Teaching to increase diversity and equity in STEM (TIDES). Retrieved January 2024, from https://www.aacu.org/initiatives/teaching-to-increase-diversity-and-equity-in-stem-tides

Association of American Universities (AAU). (2013). Framework for systemic change in undergraduate STEM teaching and learning. AAU Undergraduate STEM Education Initiative. Retrieved January 2024, from https://www.aau.edu/sites/default/files/STEM%20Scholarship/AAU_Framework.pdf

Baker, S. S., Alhassan, M. S., Asenov, K. Z., Choi, J. J., Craig, G. E., Dastidar, Z. A., ... & & Tandon, S. (2021). Students in a course-based undergraduate research experience course discovered dramatic changes in the bacterial community composition between summer and winter lake samples. *Frontiers in Microbiology*, *12*(1965), 1. https://doi.org/10.3389/fmicb.2021.579325.

Bakken, L. L., Byars-Winston, A., Gundermann, D. M., Ward, E. C., Slattery, A., King, A., ... & Taylor, R. E. (2010). Effects of an educational intervention on female biomedical scientists' research self-efficacy. *Advances in Health Sciences Education*, *15*, 167–183. https://doi.org/10.1007/s10459-009-9190-2

Bandura, A. (1997). *Self-efficacy: The Exercise of Control*. New York, NY: W.H. Freeman.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191.

Betz, N. E., & Borgen, F. H. (2000). The future of career assessment: Integrating vocational interests with self-efficacy and personal styles. *Journal of Career Assessment*, *8*, 329–338.

Bowleg, L. (2008). When black + lesbian + woman ≠ black lesbian woman: The methodological challenges of qualitative and quantitative research. *Sex Roles*, *59*(5/6), 312–325.

Castillo, W., & Babb, N. (2024). Transforming the future of quantitative educational research: A systematic review of enacting quantCrit. *Race Ethnicity and Education*, *27*(1), 1–21. https://doi.org/10.1080/13613324.2023.2248911

Castillo, W., & Gillborn, D. (2023). How to "QuantCrit:" Practices and questions for education data researchers and users. Annenberg Brown University EdWorkingPaper No. 22-546. Retrieved January 2024, from https://edworkingpapers.com/ai22-546

Catalyst. (2023, February 1). Women of Color in the United States (quick take). Catalyst. Retrieved June 2024, from https://www.catalyst.org/research/women-of-color-in-the-united-states/

Chen, C. Y., Kahanamoku, S. S., Tripati, A., Alegado, R. A., Morris, V. R., Andrade, K., & Hosbey, J. (2022). Systemic racial disparities in funding rates at the National Science Foundation. *Elife*, *11*, e83071. https://doi.org/10.7554/eLife.83071

Cobian, K., Fang, J., & Poon, O. (2022). A call for a critical intersectional lens for DEI and anti-racist strategies to include Asian Americans. Commissioned Paper for the Committee on Advancing Antiracism, Diversity, Equity, and Inclusion in STEM Organizations. National Academies of Sciences, Engineering, and Medicine. Retrieved January 2024, from https://nap.nationalacademies.org/resource/26803/ADEI_CommPaper_AsianAmerican_with_disclaimer.pdf

Cobian, K. (2019). Research self-efficacy promotes biomedical career outcomes. Literature Brief. Los Angeles, CA, Diversity Program Consortium (DPC) Coordination and Evaluation Center at UCLA. Retrieved January 2024, from https://www.diversityprogramconsortium.org/

Cobian, K. P., Zhong, S., & Guerrero, L. (2021). Examining the impact of the Building Infrastructure Leading to Diversity (BUILD) Initiative on academic and researcher self-efficacy among first-year students. *Understanding Interventions*, *12*(Suppl 1), 1–11.

Cobian, K. P., Hurtado, S., Romero, A. L., & Gutzwa, J. A. (2024). Enacting inclusive science: Culturally responsive higher education practices in science, technology, engineering, mathematics, and medicine (STEMM), *PLoS ONE*, *19*(340), e0293953. https://doi.org/10.1371/journal.pone.0293953.

Combahee River Collective. (2014). A Black feminist statement. *Women's Studies Quarterly*, *42*(3/4), 271–280. (Original work published 1977.) https://doi.org/10.1353/wsq.2014.0052

Collins, F. S., Adams, A. B., Aklin, C., Archer, T. K., Bernard, M. A., Boone, E., ... & Wolinetz, C. (2021). Affirming NIH's commitment to addressing structural racism in the biomedical research enterprise. *Cell*, *184*(12), 3075–3079.

Collins, P. H., & Bilge, S. (2016). *Intersectionality*. Cambridge, UK: John Wiley & Sons.

Collins, K. H., Price, E. F., Hanson, L., & Neaves, D. (2020). Consequences of stereotype threat and imposter syndrome: The personal journey from stem-practitioner to stem-educator for four women of color. *Taboo: The Journal of Culture and Education*, *19*(4), 10.

Covarrubias, A., & Velez, V. (2013). Critical race quantitative intersectionality: An antiracist research paradigm that refuses to "let the numbers speak for themselves." In M. Lynn& A. D. Dixson (Eds.), *Handbook of Critical Race Theory in Education* (pp. 270–285). New York, NY: Routledge.

Crenshaw, K. (1991). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, *43*, 1241–1299.

Crespi, C. M., & Cobian, K. P. (2022). A meta-analysis approach for evaluating the effectiveness of complex multisite programs. *New Directions for Evaluation*, *174*, 47–56.

Diversity Program Consortium. (n.d.) Hallmarks of success: Yr. 6-10. Retrieved January 2024, from https://diversityprogramconsortium.org/research/hallmarks

Echegoyen, L. E., Aley, S. B., Garza, J. P., Ramos, C., Oviedo, S. L., & Corral, G. (2019). Impact of open enrollment in course-based undergraduate research experiences with at-risk student populations. *EDULEARN... Proceedings*, *2019*, 6580–6588.

Forester, M., Kahn, J. H., & Hesson-McInnis, M. S. (2004). Factor structures of three measures of research self-efficacy. *Journal of Career Assessment*, *12*(1), 3–16.

Garcia, G. A., & Cuellar, M. G. (2023). Advancing "intersectional servingness" in research, practice, and policy with Hispanic-serving institutions. *AERA Open*, *9*(1), 1–8. https://doi.org/10.1177/23328584221148421

Garcia, G. A., Núñez, A. M., & Sansone, V. A. (2019). Toward a multidimensional conceptual framework for understanding "servingness" in Hispanic-serving institutions: A synthesis of the research. *Review of Educational Research*, *89*(5), 745–784.

Garcia, N. M., López, N., & Vélez, V. N. (2018). QuantCrit: Rectifying quantitative methods through critical race theory. *Race Ethnicity and Education*, *21*(2), 149–157.

Garcia, N. M., López, N., & Vélez, V. N. (Eds.). (2023). *QuantCrit: An Antiracist Quantitative Approach to Educational Inquiry*. New York, NY: Taylor & Francis.

Gelman, A., & Hill, J. (2006). Causal inference using more advanced models. In *Data Analysis Using Regression and Multilevel/Hierarchical Models* (1st ed., pp. 199–231). Cambridge: Cambridge University Press.

Ghabrial, M. A., & Ross, L. E. (2018). Representation and erasure of bisexual people of color: A content analysis of quantitative bisexual mental health

research. *Psychology of Sexual Orientation and Gender Diversity*, *5*(2), 132.

Gibbs Jr, K. D., McGready, J., & Griffin, K. (2015). Career development among American biomedical postdocs. *CBE—Life Sciences Education*, *14*(4), ar44.

Gibbs Jr, K. D., Reynolds, C., Epou, S., & Gammie, A. (2022). The funders' perspective: Lessons learned from the National Institutes of Health Diversity Program Consortium evaluation. *New Directions for Evaluation*, *2022*(174), 105–117.

Gillborn, D., Warmington, P., & Demack, S. (2018). QuantCrit: Education, policy,'Big Data'and principles for a critical race theory of statistics. *Race Ethnicity and Education*, *21*(2), 158–179.

Ginther, D. K., Kahn, S., & Schaffer, W. T. (2016). Gender, race/ethnicity, and National Institutes of Health R01 research awards: Is there evidence of a double bind for women of color? *Academic Medicine: Journal of the Association of American Medical Colleges*, *91*(8), 1098.

Goodwin, E. C., Anokhin, V., Gray, M. J., Zajic, D. E., Podrabsky, J. E., & Shortlidge, E. E. (2021). Is This Science? Students' Experiences of Failure Make a Research-Based Course Feel Authentic, *CBE—Life Sciences Education*, *20*(1076), ar10. https://doi.org/10.1187/cbe.20-07-0149.

Greifer, N. (2022). Assessing balance. *Sist Oppdatert*, *7*, 2022.

Guerard, J. J., & Hayes, S. M. (2018). Introduction to environmental chemistry of the arctic: An introductory, lab-based course offered both face-to-face and by distance. *ACS symposium series. American Chemical Society*, *1276*, 1–19. https://doi.org/10.1021/bk-2018-1276.ch001

Guerrero, L. R., Seeman, T., McCreath, H., Maccalla, N. M. G., & Norris, K. C. (2022). Understanding the context and appreciating the complexity of evaluating the Diversity Program Consortium. *New Directions for Evaluation*, *174*, 11–20.

Guo, S., & Fraser, M. W. (2014). *Propensity score analysis: Statistical methods and applications* (Vol. 11). SAGE publications.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association*, *99*(467), 609–618. https://doi.org/10.1198/016214504000000647

Katzenmeyer, C., & Lawrenz, F. (2006). National Science Foundation perspectives on the nature of STEM program evaluation. *New Directions for Evaluation*, *2006*(109), 7–18.

Kincheloe, J. L., & McLaren, P. L. (1994). Rethinking critical theory and qualitative research. In N. K. Denzin& Y. S. Lincoln (Eds.), *Handbook of Qualitative Research* (pp. 138–157). Thousand Oaks, CA: Sage.

Kozlowski, D., Larivière, V., Sugimoto, C. R., & Monroe-White, T. (2022). Intersectional inequalities in science. *Proceedings of the National Academy of Sciences*, *119*(2), e2113067119. https://doi.org/10.1073/pnas.2113067119

Leyser-Whalen, O., & Monteblanco, A. D. (2022). Course-based Undergraduate Research Experiences (CUREs) in General Education Courses. *Understanding Interventions Journal*, *13*(1), 36519.

López, N., Erwin, C., Binder, M., & Chavez, M. J. (2018). Making the invisible visible: Advancing quantitative methods in higher education using critical race theory and intersectionality. *Race Ethnicity and Education*, *21*, 180–207.

Lopez, J. D., & Gaskin, F. T. (2022). Whiteness and the erasure of indigenous perspectives in higher education. *Critical Whiteness Praxis in Higher Education* (pp. 245–255). New York: Routledge.

Maass, W., Parsons, J., Purao, S., Storey, V. C., & Woo, C. (2018). Data-driven meets theory-driven research in the era of big data: Opportunities and challenges for information systems research. *Journal of the Association for Information Systems*, *19*(12), 1. Retrieved from https://core.ac.uk/download/pdf/301378837.pdf

Maccalla, N. M. G., Gutierrez, A., Zhong, S., Wallace, S. P., McCreath, H. E., & Eagan, K. (2023). *Updated TECHNICAL REPORT: Evaluation of Postsecondary Student Outcomes: Defining Well-Represented (WRG) and Underrepresented (URG) Groups in the Diversity Program Consortium's Enhance Diversity Study using the November 2019 NIH Guidelines*. Los Angeles, CA: Diversity Program Consortium Coordination and Evaluation Center at UCLA.

Maccalla, N. M. G., Purnell, B. A., McCreath, H. E., Dennis, R. A., & Seeman, T. (2022). Gauging treatment impact: The development of exposure variables in a large-scale evaluation study. *New Directions for Evaluation*, *174*, 57–68.

Mack, K., Rankins, C., & Woodson, K. (2013). From graduate school to the STEM Workforce: An entropic approach to career identity development for STEM women of color. *New Directions for Higher Education*, *2013*(163), 23–34. https://doi.org/10.1002/he.20062.

Mack, K.M., Winter, K., & Soto, M. (Eds.) (2019). *Culturally Responsive Strategies for Reforming STEM Higher Education : Turning the TIDES on Inequity*. Bingey, UK: Emerald Publishing.

Maton, K. I., Beason, T. S., Godsay, S., Sto. Domingo, M. R., Bailey, T. C., Sun, S., & Hrabowski III, F. A. (2016). Outcomes and processes in the Meyerhoff scholars program: STEM PhD completion, sense of community, perceived program benefit, science identity, and research self-efficacy. *CBE—Life Sciences Education*, *15*(3), ar48.

McCabe, T. M., & Olimpo, J. T. (2020). Advancing metacognitive practices in experimental design: A suite of worksheet-based activities to promote reflection and discourse in laboratory contexts. *Journal of Microbiology & Biology Education*, *21*(1), 10–1128.

McCreath, H. E., Norris, K. C., Calderón, N. E., Purnell, D. L., Maccalla, N. M. G., & Seeman, T. E. (2017). Evaluating efforts to diversify the biomedical workforce: The role and function of the Coordination and Evaluation Center of the Diversity Program Consortium. *BMC Proceedings*, *11*(Suppl 12), 27. https://doi.org/10.1186/s12919-017-0087-4

McGee, E. (2018). "Black genius, Asian fail": The detriment of stereotype lift and stereotype threat in high-achieving Asian and Black STEM students. *AERA Open*, *4*(4), 2332858418816658.

McGee, E. O. (2020). Interrogating structural racism in STEM higher education. *Educational Researcher*, *49*(9), 633–644.

McGee, E. O., Main, J. B., Miles, M. L., & Cox, M. F. (2021). An intersectional approach to investigating persistence among women of color tenure-track engineering faculty. *Journal of Women and Minorities in Science and Engineering*, *27*(1), 57–84.

McGee, E. O., & Robinson, W. H. (Eds.). (2020). *Diversifying STEM: Multidisciplinary Perspectives on Race and Gender.* New Brunswick, NJ: Rutgers University Press.

Mertens, D. M. (1999). Inclusive evaluation: Implications of transformative theory for evaluation. *American Journal of Evaluation*, *20*(1), 1–14.

Mertens, D. M., & Hopson, R. K. (2006). Advancing evaluation of STEM efforts through attention to diversity and culture. *New Directions for Evaluation*, *109*, 35–51.

Miles, M. L., Agger, C. A., Roby, R. S., & Morton, T. R. (2022). Who's who: How "women of color" are (or are not) represented in STEM education research. *Science Education*, *106*(2), 229–256. https://doi.org/10.1002/sce.21694

Milner, R. H. I. V. (2007). Race, culture, and researcher positionality: Working through dangers seen, unseen, and unforeseen. *Educational Researcher*, *36*(7), 388–400.

Morris, M., & Clark, B. (2013). You want me to do WHAT? Evaluators and the pressure to misrepresent findings. *American Journal of Evaluation*, *34*(1), 57–70.

Mullikin, E. A., Bakken, L. L., & Betz, N. E. (2007). Assessing research self-efficacy in physician-scientists: The clinical research appraisal inventory. *Journal of Career Assessment*, *15*(3), 367–387.

National Institute of General Medical Sciences (NIGMS). (n.d.). *Enhancing the Diversity of the NIH-Funded Workforce. National Institute of General Medical Sciences*. Retrieved June 6, 2024, from https://www.nigms.nih.gov/training/dpc

National Institute of General Medical Sciences (NIGMS). (n.d.). *Building Infrastructure Leading to Diversity (BUILD) Initiative. National Institute of General Medical Sciences*. Retrieved June 6, 2024, from https://www.nigms.nih.gov/training/dpc/Pages/build.aspx

National Institutes of Health. (2019). Notice of NIH's Interest in Diversity. Retrieved January 2024, from https://grants.nih.gov/grants/guide/notice-files/NOT-OD-20-031.html

National Center for Science and Engineering Statistics. (2022). Survey of Graduate Students and Postdoctorates in Science and Engineering (GSS). Table 2-1. Demographic characteristics of graduate students, postdoctoral appointees, and doctorate-holding nonfaculty researchers in science, engineering, and health: 2022. Retrieved January 2024, from https://ncses.nsf.gov/surveys/graduate-students-postdoctorates-s-e/2022#data

National Research Council. (2010). *Gender Differences at Critical Transitions in the Careers of Science, Engineering, and Mathematics Faculty*. Washington, DC: The National Academies Press. https://doi.org/10.17226/12062.

National Research Council. (2012). *Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and*

Engineering. Washington, DC: The National Academies Press. https://doi.org/10.17226/13362.

Ong, M., Wright, C., Espinosa, L., & Orfield, G. (2011). Inside the double bind: A synthesis of empirical research on undergraduate and graduate women of color in science, technology, engineering, and mathematics. *Harvard Educational Review*, *81*(2), 172–209.

Ott, L. E., Godsay, S., Stolle-McAllister, K., Kowalewski, C., Maton, K. I., & La-Course, W. R. (2020). Introduction to research: A scalable, online badge implemented in conjunction with a classroom-based undergraduate research experience (CURE) that promotes students matriculation into mentored undergraduate research. *UI Journal*, *11*(1), 1–25.

Owens, E. W., Shelton, A. J., Bloom, C. M., & Cavil, J. K. (2012). The significance of HBCUs to the production of STEM graduates: Answering the call. *Educational Foundations*, *26*, 33–47.

Ovink, S. M., Byrd, W. C., Nanney, M., & Wilson, A. (2024). "Figuring out your place at a school like this:" Intersectionality and sense of belonging among STEM and non-STEM college students. *PLoS One*, *19*(1), e0296389. https://doi.org/10.1371/journal.pone.0296389

Pearson, M. I., Castle, S. D., Matz, R. L., Koester, B. P., & Byrd, W. C. (2022). Integrating critical approaches into quantitative STEM equity work. *CBE—Life Sciences Education*, *21*(1), es1.

Rios-Aguilar, C. (2014). The changing context of critical quantitative inquiry. *New Directions for Institutional Research*, *2013*(158), 95–107.

R Core Team (2022). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*. Retrieved January 2024, from https://www.R-project.org/.

Rogers, P. (2014). Overview of Impact Evaluation, methodological briefs: Impact Evaluation 1, UNICEF Office of Research, Florence.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*(1), 33–38. https://doi.org/10.2307/2683903

Sablan, J. R. (2019). Can you really measure that? Combining critical race theory and quantitative methods. *American Educational Research Journal*, *56*(1), 178–203.

Stage, F. K. (2007). Answering critical questions using quantitative data. *New Directions for Institutional Research*, *133*, 5–16.

Stage, F. K., & Wells, R. S. (2014). Critical Quantitative inquiry in context. *New Directions for Institutional Research*, *2013*(311), 1–7. https://doi.org/10.1002/ir.20041.

Stuart, E. A., & Green, K. M. (2008). Using full matching to estimate causal effects in nonexperimental studies: Examining the relationship between adolescent marijuana use and adult outcomes. *Developmental Psychology*, *44*(2), 395–406. https://doi.org/10.1037/0012-1649.44.2.395

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: A review journal of the Institute of Mathematical Statistics*, *25*(1), 1.

Suzuki, S., Morris, S. L., & Johnson, S. K. (2021). Using QuantCrit to advance an anti-racist developmental science: Applications to mixture modeling. *Journal of Adolescent Research*, *36*(5), 535–560.

Syed, M., Zurbriggen, E. L., Chemers, M. M., Goza, B. K., Bearman, S., Crosby, F. J., ... & Morgan, E. M. (2019). The role of self-efficacy and identity in mediating the effects of STEM support experiences. *Analyses of Social Issues and Public Policy*, *19*(1), 7–49.

Tabron, L. A., & Thomas, A. K. (2023). Deeper than wordplay: A systematic review of critical quantitative approaches in education research (2007–2021). *Review of Educational Research*, *93*(5), 756–786. https://doi.org/10.3102/00346543221130017

Talanquer, V. (2014). DBER and STEM education reform: Are we up to the challenge? *Journal of Research in Science Teaching*, *51*(6), 809–819.

Teranishi, R. T. (2007). Race, ethnicity, and higher education policy: The use of critical quantitative research. *New Directions for Institutional Research*, *2007*(133), 37–49.

Teranishi, R., Lok, L., & Nguyen, B. M. D. Educational Testing Service. (2013). iCount: A data quality movement for Asian Americans and Pacific Islanders in higher education. Retrieved January 2024, from https://files.eric.ed.gov/fulltext/ED573772.pdf

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. *Test Scoring*, 85–152. New York: Routledge.

Tsui, L. (2007). Effective strategies to increase diversity in STEM fields: A review of the research literature. *The Journal of Negro Education*, *76*, 555–581

United States Census Bureau. (2017). 2017 National population projections tables. Table 4: Projected race and Hispanic origin [Data set]. Retrieved July 2024, from https://www.census.gov/data/tables/2017/demo/popproj/2017-summary-tables.html

Urizar, G. G. & Miller, K. (2022). Implementation of interdisciplinary health technologies as active learning strategies in the classroom: A course redesign, *Psychology Learning & Teaching*, *21*(179), 151–161. https://doi.org/10.1177/14757257221090643.

Usher, E. L., & Pajares, F. (2008). Sources of self-efficacy in school: Critical review of the literature and future directions. *Review of Educational Research*, *78*(4), 751–796.

Wilkins-Yel, K. G., Arnold, A., Bekki, J., Natarajan, M., Bernstein, B., & Randall, A. K. (2022). "I can't push off my own mental health": Chilly STEM climates, mental health, and STEM persistence among Black, Latina, and White graduate women. *Sex Roles*, *86*(3-4), 208–232.

Wilson, D., Bates, R., Scott, E. P., Painter, S. M., & Shaffer, J. (2015). Differences in self-efficacy among women and minorities in STEM. *Journal of Women and Minorities in Science and Engineering*, *21*(1), 27–45.

Williams, J. C. (2014). Double jeopardy? An empirical study with implications for the debates over implicit bias and intersectionality. *Harvard Journal of Law & Gender*, *37*, 185.

Wong-Campbell, J. P., & Ramrakhiani, S. H. (2024). Defining mixed-race college students: Multiracial (re)categorization and the visibility of graduation gaps. *Journal of Diversity in Higher Education*. Advance online publication. https://doi.org/10.1037/dhe0000556

Zeldin, A. L., & Pajares, F. (2000). Against the odds: Self-efficacy beliefs of women in mathematical, scientific, and technological careers. *American Educational Research Journal*, *37*(1), 215–246.