Improving Measurements in Large-scale Surveys and Using Survey Data to Assess Program Impacts

Zhong, Shujin

UNIVERSITY OF CALIFORNIA
Los Angeles

Improving Measurements in Large-scale Surveys and Using Survey Data to Assess Program Impacts

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Education

by

Shujin Zhong

2021

ABSTRACT OF THE DISSERTATION

Improving Measurements in Large-scale Surveys and Using Survey Data to Assess Program Impacts

by

Shujin Zhong
Doctor of Philosophy in Education
University of California, Los Angeles, 2021
Professor Christina A. Christie, Co-Chair
Professor Minjeong Jeon, Co-Chair

Using surveys to collect data for evaluating program effectiveness is a common approach in large national multi-pronged program evaluation, and it is widely used in the evaluation of biomedical training programs initiated by the Diversity Program Consortium (DPC). The dissertation consists of three studies. In study one, *Measuring Research Mentoring Skills: Revisiting the Faculty Mentoring Competency Assessment and Developing a Short-form to Measure College Faculty-Student Mentoring*, I investigated and validated the between-item dimensionality of the Mentoring Competency Assessment (MCA) and created a short form of MCA for future evaluation that was tailored to the DPC population. In study two, *An Item Response Tree Modeling Approach for Assessing "Not Applicable" Responses in the Enhance Diversity Study*, I used an item response tree (IRTree) modeling approach for assessing "Not Applicable" or "N/A" responses, and took the measurement of faculty mentoring as an example to examine the nature of the "N/A" responses in the MCA scale and investigate within-item dimensionality. In study three, *Evaluating the Impact of the BUILD Scholar Program on First Year College Students' Intent to Pursue Science-related Research Careers*, I examined the effectiveness of the BUILD scholar program, an affiliated undergraduate diversity training program developed at

each BUILD site. I studied the influence of program participation on students' intent to pursue science-related research careers during students' initial stage in college.

The dissertation of Shujin Zhong is approved.

Mark P. Hansen

M. Kevin Eagan

Christina A. Christie, Committee Co-Chair

Minjeong Jeon, Committee Co-Chair

University of California, Los Angeles

2021

To my dear mom and dad

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

It was a privilege to work with so many wonderful scholars. First, I would like to express tremendous gratitude to my advisor and committee co-chair, Dr. Tina Christie, who helped me stay focused and motivated and provided steadfast support and resources throughout my doctoral studies. Many thanks to my dissertation co-chair, Dr. Minjeong Jeon, a role-model researcher and top-tier psychometrician, who inspired me to dive into the field of measurement and guided my dissertation work. I would also like to acknowledge my dissertation committee members, Dr. Mark Hansen and Dr. Kevin Eagan, for their invaluable feedback. Mark's courses fundamentally supported my dissertation analysis, and Kevin's insights into research in higher education compelled me to deepen and improve my analyses. A special thank you goes to Dr. Steve Wallace, my former committee member and research-team group leader. I learned so much from Steve's keen feedback, line-editing, and suggestions for professional development. Although Steve is no longer with us, his dedication to research and mentoring over the course of his career continues to inspire me.

The past five years at UCLA have been a memorable journey that all began in a research method class during the summer of 2015, when Dr. Nicole M.G. MacCalla encouraged me to pursue a doctoral program. In that moment I remember thinking that if doctoral training could help me learn as much as she knew about research, I would do whatever it took to get a PhD. Nicky introduced me to the Social Research Methodology (SRM) program and the UCLA Coordination and Evaluation Center (CEC), where I would study and train. I would like to thank Nicky for seeing something in me, believing in me, and supporting me beyond academic and research training. She has been a great mentor, a reliable colleague, and my co-author in chapters two and four, as well as many other publications. I would also like to thank her for bringing me into her family; since then, the Gerardi family has stood by me, sharing my joys, laughter, and tears.

If I were asked to choose again, I would still choose the SRM program. Thanks to the following SRM faculty members (in addition to those on my committee): Drs. Marv Alkin, Li Cai, Ananda Marin, Felipe Martinez, Teresa McCarty, Mike Rose, Mike Seltzer, and Reenie Webb. They sparked me intellectually, helped me grow academically, and supported my extensive research interests. I miss those good old days in their classes and RACs, and I enjoyed every intellectual conversation. They helped me realize and accept that in methodology, "nothing is [absolutely] true; everything is permitted." I would like to extend a special thank you to Prof. Mike Rose, who told me that I was a good writer and encouraged me to write freely. Mike will be missed as a great teacher and an excellent writer. I kept Mike's book Writer's Block: The Cognitive Dimension by my side as I wrote this dissertation, constantly trying to figure out my own writer's block. The first sentence in Mike's book is a quote from Gustave Flaubert, "You don't know what it is to stay a whole day with your head in your hands trying to squeeze your unfortunate brain so as to find a word" – that describes some of my toughest days during the past two years. I would not have survived those days without support from my peers, my fellow SRMers – especially those from the evaluation group and Tina's RAC, Maria-Paz Fernandez, Christine Liboon, Lindy Messer, Nadia Sabat Bass, Evelyn Wang, Dr. Emi Fujita-Conrads, and Dr. Kristen Rohanna; my cohort, Yonsoo Suh, Junok Kim, Dr. Sijia Huang, and Dr. Mariana Barragan; my classmates and study buddies, Jinwen Luo, Zhaopeng Ding, Tom Jacobson, and Dr. Seungwon Chung; and so many others who have encouraged me throughout my journey. I would especially like to thank Maria-Paz, a dear friend and great listener, for her support and final push during the past few months. In addition, thanks to the capable team in the Office of Student Services, especially, Dr. Amy Gershon, who is brilliant and compassionate and knows everything about student-service procedures and regulations, literally.

I am grateful I had the opportunity to work with the CEC for six years on the Enhance Diversity Study. The CEC housed the data I used for my dissertation and generously shared their resources, including some of the figures and tables appearing in my dissertation. I would like to thank the PIs, Dr. Keith Norris and Dr. Teresa Seeman, for leading the CEC and for their constant support of my work. I extend my gratitude to Dr. Krystle Cobian and Dr. Kate Crespi, colleagues in the analysis group, for their efforts to protect my time and allow me to focus on my priorities. Thanks to colleagues in the Evaluation Core: Gerald Bempong, Dr. Lourdes Guerrero, Weijuan Han, Dr. Sylvia Hurtado, Dr. Cynthia Joseph, Dr. Heather McCreath, Terry Nakazono, Dawn Purnell, Karina Ramirez, Dr. Hector Ramos, Ana Romero, and other colleagues in the CEC, for their feedback, insights, and assistance.

Finally, I would like to thank my dear parents for their endless love, measureless support, and unconditional trust. They planted the seed of learning in me, which grew into my lifelong passion. They encouraged me to ask questions and seek the truth when I started elementary school, and I would carry over these suggestions to my future academic career. Ba ba ma ma, xie xie ni men!

VITA

EDUCATION

2013	Bachelor of Arts, English Huazhong University of Science and Technology, China
2013	Bachelor of Management, International Business Huazhong University of Science and Technology, China
2016	Master of Education, Postsecondary Administration and Student Affairs University of Southern California

WORK

2016-2021	Graduate Student Researcher Diversity Program Consortium, Coordination and Evaluation Center, University of California, Los Angeles
2019-2020	Teaching Assistant Department of Sociology, University of California, Los Angeles
2020	Teaching Associate Department of Sociology, University of California, Los Angeles

CHAPTER 1

Introduction of the Diversity Program Consortium

1.1 Background

The US is becoming more racially and ethnically diverse, and this trend will continue to grow in the coming decades (Cohn & Caumont, 2016). In sharp contrast to the population's growing diversity is the stagnant demography of the US biomedical research workforce, which lags behind in participation of workers from historically excluded groups despite numerous efforts over the past 40 years to increase diversity (Valantine & Collins, 2015; McGee Jr, Saran, & Krulwich, 2012). Individuals from historically marginalized groups in science and research (defined in Maccalla, Gutierrez, Zhong, Wallace, & McCreath, 2020) are ”disproportionately underrepresented and underserved at all levels of the scientific workforce” (Cobian and Gutiérrez, 2021, p. 3), from undergraduate students and their pipelines, to faculty members. The completion of biomedical degrees — at the undergraduate and graduate levels — by underrepresented groups (URGs; Maccalla et al., 2020) has continuously fallen behind (Rask, 2010; Valantine & Collins, 2015) that of their well-represented peers (WRGs; Maccalla et al., 2020). Further, scientists who belong to the URGs are significantly less likely to be awarded research grants from the National Institutes of Health (NIH) compared to their WRG counterparts (Ginther et al., 2011).

In response to these continued disparities in the biomedical workforce, the NIH funded a new set of initiatives in 2013. The aim was unique in that it employed a broad and transformative approach to “[promote] diversity in the NIH-funded biomedical, behavioral, clinical, and social sciences (collectively termed ‘biomedical’) research workforce" (Funding Opportunity Announcement, 2013a, 2013b, 2013c), providing constant support for individuals from diverse backgrounds underrepresented in biomedical research. Participants receive training and mentoring, from as early as their undergraduate studies, through their terminal degrees, and into their early career after achieving their terminal degrees. By providing training and mentoring, the program aims to contribute to diversifying the candidate pool in biomedical research at different educational stages. The funding opportunity resulted in the creation of Enhancing the Diversity of the NIH-Funded Workforce (n.d.), also known as the Diversity Program Consortium (DPC), a trans-NIH program managed by the National Institute of General Medical Sciences (NIGMS).

1.2 Diversity Program Consortium (DPC)

The DPC consists of three core integrated initiatives: the Building Infrastructure Leading to Diversity (BUILD) Initiative, the National Research Mentoring Network (NRMN) Initiative, and the Coordination and Evaluation Center (CEC). The DPC develops, implements, assesses, and disseminates innovative and effective approaches to research training and mentoring, with the goals of: "1) engaging, training and mentoring students 2) enhancing faculty development, and 3) strengthening institutional research training infrastructure" (Enhancing the Diversity of the NIH-Funded Workforce, n.d.). The DPC emphasizes impacts on students, faculty, and institutions, at all levels of the biomedical workforce. Each of the three components (i.e., BUILD, NRMN, and CEC) of the DPC provide a wide range of support for URGs in biomedical research fields.

1.2.1 BUilding Infrastructure Leading to Diversity (BUILD)

The disparity in diversity between the general population and that of biomedical research field professionals can be reduced only if the candidate pool for the field consists of a diverse population. In 2014, the NIH initiated the BUILD program in order to foster interest among underrepresented undergraduates from diverse backgrounds in biomedical research fields (as broadly defined in the Funding Opportunity Announcement, to include biomedical, behavioral, clinical, and social sciences, 2013b). The BUILD initiative, a set of quasi-experimental training programs at 10 primary BUILD institutions across the nation, aims to attract students from diverse and underserved backgrounds into biomedical research fields and prepare them for academic success and career readiness through innovative methods (McCreath et al., 2017). Ten institutions are currently funded as primary BUILD sites (Figure 1.1, cited from Davidson et al., 2017, p. 158). Table 1.1 (adapted from Davidson et al., 2017, p. 166) summarized the basic institutional information of the BUILD primary sites prior to the start of the BUILD program. These primary BUILD sites, along with their research partner institutions, provide biomedical training and mentoring to undergraduate students in the BUILD programs.

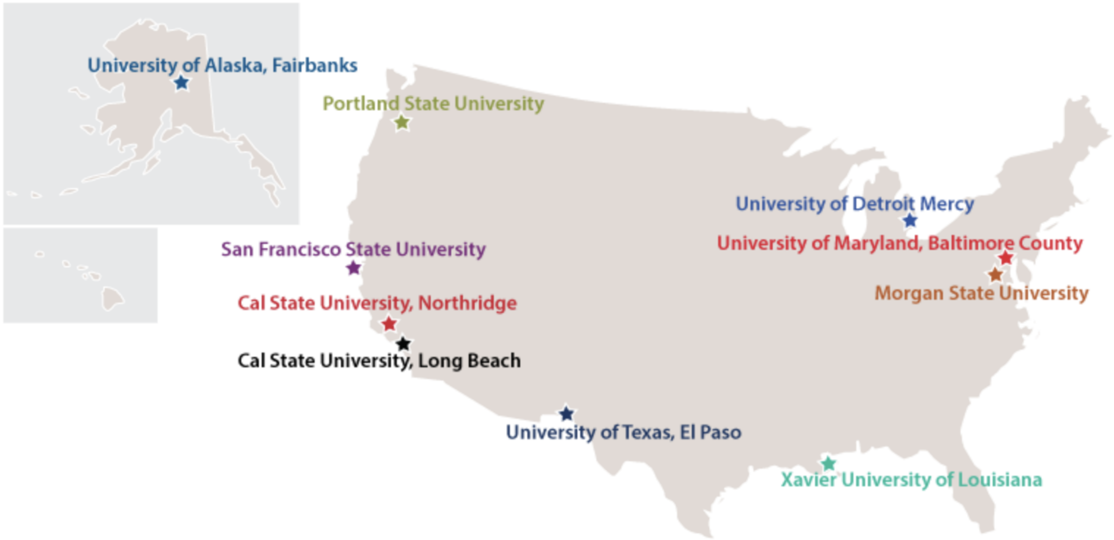

University of Alaska, Fairbanks
Portland State University
University of Detroit Mercy
University of Maryland, Baltimore County
San Francisco State University
Morgan State University
Cal State University, Northridge
Cal State University, Long Beach
University of Texas, El Paso
Xavier University of Louisiana

Figure 1.1: BUilding Infrastructure Leading to Diversity (BUILD) Primary Sites

Note. From ”A participatory approach to evaluating a national training and institutional change initiative: the build longitudinal evaluation,” by P. L. Davidson, N. M. G. Maccalla, A. A. Afifi, L. Guerrero, T. T. Nakazono, S. Zhong and S. P. Wallace, 2017, *BMC proceedings (Vol. 11)* p. 158. Copyright 2017 by the author(s). Reprinted with permission.

Table 1.1: BUILD Prime Sites: Institutional Characteristics

Site	Type	Ave. total NIH fund. 2011-13	Adm. rate	Ave. SAT	Pell Grants (%)	6-yr Grad. rates	Total students	URM (%)	Incoming 1st-yr	Transfer students (%)	Grad. students (%)
CSULB	Public	$4.3M	35%	1055	50%	57%	36809	42%	4335	12%	14%
CSUN	Public	$4.7M	53%	915	52%	48%	40131	48%	5526	16%	11%
SFSU	Public	$7.9M	66%	990	43%	47%	29465	29%	3754	12%	12%
PSU	Public	$5.2M	69%	1030	41%	42%	27696	15%	1703	15%	20%
UAF	Public	$8.8M	-	-	29%	33%	8620	20%	942	6%	13%
UTEP	Public	$11.3M	100%	-	58%	38%	23079	85%	3256	10%	14%
XULA	HBCU	$5.0M	66%	1005	53%	47%	2976	79%	579	5%	21%
UDM	Private	$0.0M	69%	1121	29%	57%	4945	14%	469	7%	44%
UMBC	Public	$9.0M	60%	1210	28%	61%	13979	22%	1629	11%	19%
MSU	HBCU	$1.6M	65%	880	62%	29%	7698	87%	1078	7%	6%

Note. Original information was pulled from the Institute of Education Sciences (IES) National Center for Education Statistics (NCES) - Final release data, 2013-2016 (https://nces.ed.gov/ipeds/datacenter/). Adapted from ”A participatory approach to evaluating a national training and institutional change initiative: the build longitudinal evaluation,” by P. L. Davidson, N. M. G. Maccalla, A. A. Afifi, L. Guerrero, T. T. Nakazono, S. Zhong and S. P. Wallace, 2017, *BMC proceedings (Vol. 11)* p. 164. Copyright 2017 by the author(s). Adapted with permission. See detailed explanation in the original Davidson et al. (2017) article.

1.2.2 National Research Mentoring Network (NRMN)

The NRMN initiative (Figure 1.2, cited from NRMN, 2019) is a nationwide consortium of more than 100 partner institutions and organizations that provides networking, mentorship, and training opportunities for researchers from diverse backgrounds. The NRMN aims to develop, implement and disseminate innovative, evidence-based best practices to improve mentoring relationships in biomedical research fields across the consortium (Funding Opportunity Announcement, 2013c; NRMN, n.d.). Unlike the BUILD programs, which serve the undergraduate participants and their faculty mentors from the 10 primary sites, the NRMN programs recruit participants at all levels (from undergraduate students to early-career faculty) and work with partner institutions and organizations across the United States (NRMN, n.d.).

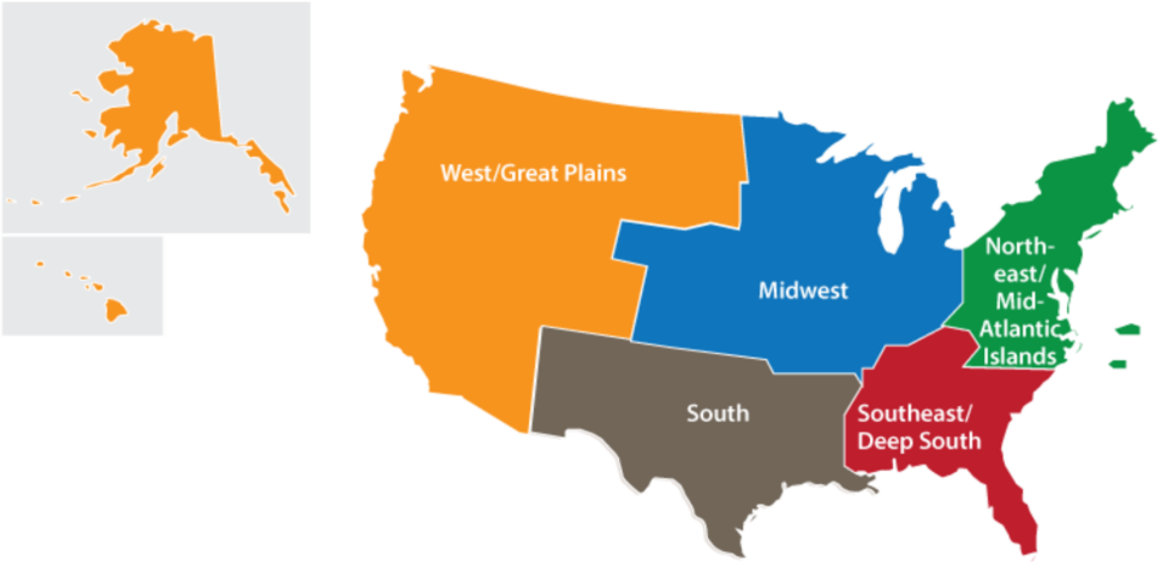


Figure 1.2: National Research Mentoring Network (NRMN)

Note. From National Research Mentoring Network (NRMN), by Diversity Program Consortium (DPC), 2019 https://www.diversityprogramconsortium.org/pages/nrmn). In the public domain. Reprinted with permission.

1.2.3 Coordination and Evaluation Center (CEC)

The Coordination and Evaluation Center (CEC) has the responsibility of evaluating and assessing the effectiveness of the BUILD and NRMN initiatives. The key foci for the CEC evaluation of the DPC programs include but are not limited to 1) developing and revising the DPC "Hallmarks of Success" (2020) in biomedical research career paths, 2) exploring motivations and factors that contribute to student participation in biomedical research career paths, 3) identifying institutional, social, and individual factors that influence students' decision to pursue biomedical careers, 4) recognizing institutional structures and re- sources that support student success, and 5) finding approaches to continuing impact beyond the funding period (Funding Opportunity Announcement, 2013a).

Aligned with the Funding Opportunity Announcement (2013a), the CEC, along with the other DPC initiatives, identified the "important indicators of transition through [biomedical] career stages" (McCreath et al., 2017, p.16), and grouped and named them as DPC Hallmarks of Success. At the beginning of the BUILD program evaluation, the Executive Steering Committee (ESC, composed of BUILD and NRMN PIs as well as representatives from NIH and the CEC) developed a checklist of potential indicators, based on the literature and program implementation. After discussion, the ESC voted on each indicator then considered the indicators that received at least 80% approval to be Hallmarks. The Hallmarks are indicators of transition from entering into biomedical fields through various career stages. Thus, growth in a Hallmark could reflect the increasing odds of success in the biomedical field. The Hallmarks of Success (2020) are presented on the DPC website, and the Hallmarks have been mapped to the DPC evaluation logic models.

To understand the impact of these multi-pronged research training activities on future career success in the biomedical field, the CEC identified and summarized the DPC Hallmarks as outputs and outcomes of the DPC programs. The Hallmarks of Success define the constructs of interests to be measured using the survey items, and these constructs are indicators of individual development or indicators of successful pursuit of biomedical education and career paths. To assess program effectiveness, the CEC needs to assess the participants' growth and achievement of the Hallmarks throughout the programs. Accordingly, the CEC developed a longitudinal multi-method evaluation plan that includes utilizing large-scale surveys to assess stakeholder credibility (internal reliability) and scientific credibility (external reliability).

1.3 Organization of Research

Using surveys to collect data for evaluating program effectiveness is a common approach in large national multi-pronged program evaluation, and it is widely used in the evaluation of biomedical training programs initiated by the DPC. In this dissertation, I utilized the survey data and program participation data from the DPC evaluation to conduct three studies.

1.3.1 Study One: Developing a Short-form to Measure Faculty Mentoring

In study one, *Measuring Research Mentoring Skills: Revisiting the Faculty Mentoring Competency Assessment and Developing a Short-form to Measure College Faculty-Student Mentoring*, I investigated and validated the between-item dimensionality of the Mentoring Competency Assessment (MCA Fleming et al., 2013) and created a short form of MCA for future evaluation that was tailored to the DPC population. To measure the Hallmarks that were indicated by latent traits, the CEC identified existing scales as item pools and tailored the scales to the DPC population. Although most existing scales were assessed when they were created, the scales were not purposefully designed based on the DPC population nor for the DPC evaluation. Using existing scales without performing any validating test on the DPC population could be problematic, especially for a national high-stake multi-site longitudinal study. I removed problematic items and selected suitable items based on measurement properties, and created the MCA-short-C, a short form of MCA scale for measuring college faculty-student research mentoring. The paper provided supportive evidence demonstrating that the 9-item MCA-short-C kept the features of the original 26-item MCA scale, and measured the college faculty-student research mentoring adequately well for both faculty (self-rating) and student (rating mentors) population.

1.3.2 Study Two: Using IRTree Models to Assess "N/A" Responses

In study two, *An Item Response Tree Modeling Approach for Assessing "Not Applicable" Responses in the Enhance Diversity Study*, I used an item response tree (IRTree) modeling approach for assessing "Not Applicable" or "N/A" responses, and took the measurement of faculty mentoring as an example to examine the nature of the "N/A" responses in the MCA scale and investigate within-item dimensionality. The DPC surveys provided the "N/A" option as a response category so participants could have more opportunities to express their actual conditions. However, how to interpret "N/A" responses has rarely been studied. These responses were often treated by analytical models as missing, although by design, participants were provided opportunities to distinguish the use of "N/A" from a missing response. This set an example of handling "N/A" responses, which could be a reference for DPC data analysis and for interpreting other similar response options.

1.3.3 Study Three: Evaluating the Impact of the BUILD Scholar Program

The first two studies aimed at addressing measurement issues. After using a valid scale to measure the construct of interest and collecting the data from the DPC population, I was able to evaluate the impacts of the DPC programs. In study three, *Evaluating the Impact of the BUILD Scholar Program on First Year College Students' Intent to Pursue Science-related Research Careers*, I examined the effectiveness of the BUILD scholar program, an affiliated undergraduate diversity training program developed at each BUILD site. I studied the influence of program participation on students' intent to pursue science-related research careers during students' initial stage in college. In this study, I also demonstrated potential approaches, such as matching, regression analysis, and sensitivity analysis for assessing program impacts using survey data. The results indicated that the BUILD scholar program positively influenced students' intent to pursue science-related research careers during students' initial stage in college.

The three studies in this dissertation utilized the DPC survey data to address important issues in measurement, survey methods, and evaluation of program impacts. Using surveys to collect data for evaluating program effectiveness is a common approach in large national multi-pronged program evaluation. These three studies demonstrated approaches for improving measurement and in large-scale surveys and using survey data to assess program impacts.

CHAPTER 2

Measuring Research Mentoring Skills: Revisiting the Faculty Mentoring Competency Assessment and Developing a Short-form to Measure College Faculty-Student Mentoring

2.1 Introduction

2.1.1 Enhance Diversity Study

Under the Diversity Program Consortium (DPC), a trans-NIH program managed by the National Institute of General Medical Sciences (NIGMS), the Building Infrastructure Leading to Diversity (BUILD) initiatives and the National Research Mentoring Network (NRMN) initiatives provide research training and mentoring programs for individuals from historically marginalized groups in science and research. The Enhance Diversity Study (the evaluation of the DPC), supported by the National Institutes of Health (NIH), is determining the effectiveness of innovative approaches to engage individuals from diverse backgrounds and help them prepare for and succeed in biomedical research careers (Davidson et al., 2017; McCreath et al., 2017).

The DPC supports transformative approaches to student engagement, research training, mentoring, faculty development, and infrastructure development. The Enhance Diversity Study administers annual surveys to students and faculty at 10 primary BUILD institutions (2015-2024) as well as mentees and mentors in the NRMN (2015-2019) programs, with the intention of measuring key outcomes of interest, i.e., the Hallmarks of

Success (Davidson et al., 2017; McCreath et al., 2017).

2.1.2 Faculty Mentoring in the Diversity Program Consortium

Early in the project, DPC leadership adopted a set of measurable objectives dubbed the "Hallmarks of Success" (or "Hallmarks"). These Hallmarks represent key indicators of students', faculty members', and institutions' ability to meet the kinds of benchmarks identified in previous literature as relevant to individual and organizational success with respect to advancement and achievement within STEM education and within the STEM workforce. These Hallmarks cover a wide-range of aspects, such as quantity of mentoring, quality of mentoring, improving mentoring skills and mentoring with diverse minds (Hallmarks of Success, 2020). In the current version of the Hallmarks of Success (2020), three (out of 18) student Hallmarks, four (out of 17) faculty Hallmarks and one (out of 11) institutional Hallmark are directly related to mentoring (Table 2.1). Preparation for and mentoring of undergraduate and graduate students, post-docs, and junior faculty compose a key area of focus for the DPC's activities.

It is expected that students participate in mentored research, receive frequent mentoring, and are satisfied with the quality of mentorship. Faculty are expected to engage in mentor training, mentor frequently, and mentor with high self-efficacy while utilizing evidence-based practices. Institutions (BUILD primary sites) are expected to demonstrate commitments to implementing and sustaining mentoring practices to enhance diversity in the biomedical research workforce. Additionally, the major focus of the NRMN is to develop, implement and disseminate innovative, evidence-based best practices to improve research mentoring relationships at all levels (from undergraduate students to early-career faculty) in the biomedical research field (Funding Opportunity Announcement, 2013c; NRMN, n.d.).

The Hallmarks of Success affirms the essentials of faculty mentoring, and particularly, they indicate that mentoring matters both ways. It is important for the DPC to help faculty improve their mentoring competency, and it is equally important to know whether students are satisfied with the research mentoring they receive.

Table 2.1: Mentoring Related Hallmarks

Types	ID	Hallmarks
Student	STU-4	Satisfaction with quality of mentorship
	STU-10	Frequent receipt of mentoring to enhance success in the biomedical pathway
	STU-11	Participation in mentored or supervised biomedical research
Faculty	FAC-3	High self-efficacy as a mentor to biomedical research trainees
	FAC-4	High self-efficacy as a mentor to a diverse group of biomedical research trainees
	FAC-5	Frequently mentors students, post-docs, and/or more junior faculty on biomedical related issues
	FAC-17	Uses evidence-based practices in teaching and mentoring
Institutional	INST-10	Demonstrated institutional commitment to implementing and sustaining mentoring practices that promote the development of research-oriented students from all backgrounds

2.1.3 Purpose of the Study

According to prior research findings on the significance of quality mentorship in supporting the academic and career progression of students from underrepresented backgrounds, quality measurement tools are of great importance in capturing impacts of training programs. We are unable to assess the growth of a latent trait, unless we can express the quantity (Thomson, 1889). Before we evaluate DPC programs' effectiveness at improving faculty mentoring in research or attempt to measure students' satisfaction with research mentoring, we need to find tools to measure faculty-mentoring competency. Using survey data to evaluate program effectiveness hinges upon having valid and reliable measurement tools. A common approach is to identify in literature existing valid measures of the construct of interests. In this process, we have to ensure that the identified scales and survey items are measuring the same population in the original scale development and in our program.

To collect reliable information on faculty mentoring competency for the NRMN and the BUILD program participants, we incorporated the 26-item Mentoring Competency Assessment (MCA) scale, originally developed by Fleming et al. (2013), into the DPC surveys. One feature of many large-scale studies, including the Enhance Diversity Study, is the measurement of multiple outcomes of interest. When scales include a multitude of items, surveys can quickly become unwieldy in length and pursuing item reduction procedures becomes necessary. From a practical consideration, since survey length and time to completion are negatively correlated with response rates and survey completion, having the fewest number of items as are necessary on a survey is of paramount importance.

Although the MCA provides a good item bank for measuring faculty mentoring skills, the 26-items take up substantial space on the Enhance Diversity Study student and faculty surveys where many other important constructs are also being measured. An unnecessarily long survey jeopardizes survey completion. In an effort to eliminate item redundancy and reduce respondent burden, item reduction procedures are pursued. The most informative items of the scale are identified to produce a short unidimensional scale measuring faculty mentoring competency. In this study, our goal is to assess and increase the feasibility and effectiveness of using the MCA scale to measure mentoring competency in large scale surveys. Item reduction is built upon confirming the measurement validity on the DPC population, with reduced scales maintaining the overall integrity of the original MCA scale. This paper details the results from item reduction procedures applied to the original long form (26-items) of the MCA (Fleming et al., 2013).

The previously established short form of the Mentoring Competency Assessment, the MCA-short (Zhong, Maccalla, & Jeon, 2020), largely consulted the NRMN survey responses. The NRMN programs recruited participants at all levels, from undergraduate students to early-career faculty, (NRMN, n.d.); however, the majority of their mentor-mentee relationships were not at the undergraduate level. As tested in the first MCA-short development (Zhong et al., 2020) practice, the item context of the original MCA scale fit the NRMN population better, due to the similarity of the survey participants (Graduate level academic research or clinical research mentor-mentee pairs).

Another purpose of this study is to establish a short form of the MCA that allows cross-rating from both faculty mentors and student mentees of faculty research mentoring competency for faculty-student research mentoring at the undergraduate level (namely MCA-short-College or MCA-short-C), while maintaining similar psychometric properties of the original scale. It is important that the high validity and reliability of the original MCA scale transfer to the MCA-short-C. Getting assessment of mentoring skills in college settings down to a reasonable length on the Enhance Diversity Study surveys supports complete and sustained engagement in the longitudinal study of DPC effectiveness. In this study we take an exploratory approach to examine how well the MCA measures faculty mentoring of the BUILD faculty and students, and to investigate the extent to which we can shorten the MCA to a reasonable length while measuring the faculty mentoring competency construct reliably.

2.2 Literature Review

2.2.1 Faculty Mentoring

In Homer's Odyssey, "mentor" was a "wise and trusted counselor" whom Odysseus entrusted with the care and education of his son, Telemachus. Along these lines, Crisp and Cruz (2009) defined a mentor as "a wise, responsible and trusted advisor" who guides the development of an individual (p. 527). Previous studies have concluded that mentorship programs provide beneficial outcomes for individuals and aid in the development of organizations or institutions (Lin & Hsu, 2012; Seibert, 1999).

The Faculty-student mentoring model is the most common form of mentorship in college (Zhong, 2016), and the effectiveness of faculty mentoring has been considered as one of the important indicators of institutional change (Bradbury & Koballa Jr, 2008; Colvin & Ashman, 2010; Lin & Hsu, 2012; Seibert, 1999). However, our understanding of which specific aspects of mentoring lead to students' career advancement and improved psychosocial support is largely anecdotal (Crisp & Cruz, 2009; Gómez, Ali, & Casillas, 2014). Jacobi (1991) as well as Crisp and Cruz (2009) holistically reviewed studies on mentorship in higher education, and recognized the limitations of these studies: limited sample sizes, a lack of evidence from quasi-experimental designs, and measurement issues that led to low external validity.

In 2010, five awardee institutions of the NIH Clinical and Translational Science Awards (CTSA) jointly developed a clinical and translational research mentor training curriculum, based on the Entering Mentoring seminar, a mentor training program in the science, technology, engineering, and mathematics (STEM) fields (Pfund et al., 2013, 2014). This training program featured a large, experimental nation-wide study that involved participants from 16 U.S. universities. This experimental program had a fairly large sample size of 283 pairs of mentors and mentees, and more importantly, the program provided evidence of rather high external validity. Having overcome the previously mentioned limitations of past studies, the CTSA training program and the evaluation of this program remains a vital addition to mentorship literature.

2.2.2 The Mentoring Competency Assessment (MCA) Scale

Along with the CTSA program, Fleming et al. (2013) developed and utilized the Mentoring Competency Assessment (MCA) scale for evaluating the mentor training, to fulfill the requirement by the CTSA (Pfund et al., 2013, 2014). The original MCA scale went through a rigorous development process including 1) a holistic review of the Mentorship Effectiveness Scale (a 12-item six-point agree–disagree-format Likert-type rating scale developed by Berk, Berg, Mortimer, Walton-Moss, & Yeo, 2005) and other instruments, 2) interviews with mentoring program participants, and 3) scale validation procedures on the CTSA participants – 283 pairs of mentors (self-rating) and mentees (rating their mentors) at 16 U.S. universities using confirmatory factor analysis (CFA).

Fleming et al. (2013) identified 26-item MCA scale that measured six sub-domains of mentoring competency: 1) maintaining effective communication, 2) aligning expectations, 3) assessing understanding, 4) addressing diversity, 5) fostering independence, and 6) promoting professional development. Table 2.2 presented the six domains, their related mentoring skill names, and the items that measured the skills. These 26 items asked participants (both mentors and mentees) to rate the mentors' mentoring competence on a 7-point Likert-type scale, ranging from not at all skilled (1) to extremely skilled (7). The scale was then utilized to estimate the improvement of effective mentoring in the CTSA program (Pfund et al., 2014). The articles by Pfund et al. (2013, 2014) and Fleming et al. (2013) promoted various implementations in studies and evaluations of mentorship and mentor training.

2.2.3 The Development of the MCA-short Scale in the Enhance Diversity Study

The 26-item MCA scale was used to evaluate faculty research mentoring competency in the national longitudinal Enhance Diversity Study, with permission from its original developers. In an effort to reduce respondent burden in the Enhance Diversity Study, analysts examined the psychometric properties of the MCA, using faculty/mentor data from two biomedical faculty survey samples, the BUILD Faculty Annual Follow-up Survey (FAFS) 2017-2018 sample and the NRMN Follow-up Survey 2016-2018. Statistical analyses employed to test the quality of each item and the overall scale included Exploratory Factor Analysis (EFA), Confirmatory Factor Analysis (CFA), and Item Response Theory (IRT). The results indicated that the MCA scale was reliable for measuring the mentors self-assessment of their research mentoring skills. After analyzing the psychometric properties of items in the MCA scale, analysts reduced the original MCA scale to a short form.

Table 2.2: The Mentoring Competency Assessment (MCA) Scale

Sub-domains	Skills	Items
Communication	Listening	Active listening
	Feedback	Providing constructive feedback
	Trust	Establishing a relationship based on trust
	Styles	Identifying and accommodating different communication styles
	Strategies (C)	Employing strategies to improve communication with mentees
	Coordinate	Coordinating effectively with mentees' other mentor
Expectation	Set expectations	Working with mentees to set clear expectations of the mentoring relationship
	Align expectations	Aligning expectations with mentees'
	Differences	Considering how personal and professional differences may impact expectations
	Goals	Working with mentees to set research goals
	Strategies (E)	Helping mentees to develop strategies to meet goals
Assessing	Knowledge	Accurately estimating mentees' level of scientific knowledge
	Mentee ability	Accurately estimating mentees' ability to conduct research
	Mentee skills	Employing strategies to enhance mentees' knowledge and abilities
Independence	Motivation	Motivating mentees
	Confidence	Building mentees' confidence
	Creativity	Stimulating mentees' creativity
	Contributions	Acknowledging mentees' professional contributions
	Negotiating	Negotiating a path to professional independence with mentees
Diversity	Prejudice	Taking into account the biases and prejudices the mentor brings to the mentor/mentee relationship
	Background	Working effectively with mentees whose personal background is different from the mentor (age, race, gender, class, region, culture, religion, family composition etc.)
Profession	Network	Helping mentees network effectively
	Career goals	Helping mentees set career goals
	Work/life balance	Helping mentees balance work with their personal life
	Role model	Understanding the mentor's impact as a role model
	Acquire resources	Helping mentees acquire resources (e.g. grants, etc.)

Rating on a 7-point Likert-type scale: 1 = not at all skilled, 4 = moderately skilled, 7 = extremely skilled

The resulting short form of the Mentoring Competency Assessment (MCA-short; Zhong et al., 2020) included 8 items measuring faculty mentoring competency across the 6 sub-domains (Table 2.3). IRT scores from the 8-item scale and the 26-item scale were highly correlated (above .96 for both BUILD faculty and NRMN mentor samples). The MCA-short was proved to be both valid and reliable on the mentor population and was offered as an alternative to the long form in measuring faculty mentoring competency, particularly when researchers were concerned about space constraints and/or respondent burden. The MCA-short was used in the subsequent Enhance Diversity faculty/mentor surveys and studies associated with the DPC. Permission from original MCA developers was granted to publish the MCA-short.

Table 2.3: The Short Form of the Mentoring Competency Assessment (MCA-short)

Sub-domains	**Items**
Communication	Establishing a relationship based on trust
Expectation	Aligning the mentor's expectations with mentees'
Assessing	Accurately estimating mentees' level of scientific knowledge
Independence	Building mentees' confidence Stimulating mentees' creativity
Diversity	Taking into account the biases and prejudices the mentor brings to mentor/mentee relationship
Profession	Helping mentees balance work with personal life Understanding mentor's impact as a role model

Rating on a 7-point Likert-type scale: 1 = not at all skilled, 4 = moderately skilled, 7 = extremely skilled

2.3 Methods

Our study examined the generalizability and transferability of the MCA scale in the Enhance Diversity Study, and explored variations of the MCA scale's appearance in large-scale surveys in higher education. In Fleming et al. (2013), items in the MCA scale were selected based on both mentors' self-rating as well as mentees' rating on their mentors. To expand the usage and coverage of the short form of the MCA scale on the DPC population, in this paper, we presented the item selection process using the BUILD faculty responses and the cross-validation with the corresponding mentee responses. The purpose of this study is to re-establish the MCA-short from the faculty-student samples in college settings to create a new short form, the MCA-short-C, while maintaining similar psychometric properties of the original scale. Sequential statistical analyses were employed to test the quality of each item and the overall scale included factor analysis, item response theory and cross-validation procedures (mentor/mentee ratings).

2.3.1 Data and Sample

In the DPC administration of the MCA scale, we strictly followed the original design of the MCA scale. We asked the faculty/mentor participants to rate their mentoring skills on a 7-point Likert-scale from 1-7, representing "not at all skilled" to "extremely skilled" of the ability the items described. We used the same questions in the student/mentee survey, and asked students to rate the skill level of their primary mentor on the same scale. In addition to the 7 options, we also provided an extra "N/A" option, so that participants did not have to respond to the items that were not applicable to their experience. For items that we kept for analyses, we temporarily treated the "N/A" responses as missing data at random.

The data utilized in this paper were faculty/mentor and student/mentee responses to the MCA scale from the BUILD Faculty Annual Follow-up Survey (FAFS, 2017-18) and the BUILD Student Annual Follow-up Survey (SAFS, 2017). The BUILD faculty were from the 10 primary BUILD sites. Most of the 10 BUILD sites were teaching universities with large percentages of undergraduate students. The BUILD FAFS was distributed to faculty members in the primary BUILD sites, and we received 683 responses in the survey year 2017-18. Among them, 586 participants self-identified as a mentor, and 547 mentors responded on the scale of 1-7 to at least half of the questions in the MCA scale. Within those 547 mentors, 312 faculty respondents identified their mentees through training programs or departmental assignments. The SAFS was distributed to students in the BUILD sites, and we received 5230 responses in 2017. Among them, 1482 participants self-identified as having a faculty mentor and responded on the scale of 1-7 to at least half of the questions in the MCA scale. Within these 1482 students, 851 of them identified their mentors through training programs or departmental assignments, including 353 students who found their faculty mentors through the BUILD programs. In this study, we used the 547 faculty responses and 1482 student responses to at least half of items in the MCA scale, because we wanted to ensure that after item reduction when we dropped more than half of items in the MCA, the responses were from the same groups of respondents.

We followed the procedure proposed by the scale developer, and validated the scale using data from both mentors' self-rating and mentees' rating of their mentors. Unlike in the original scale development article by Fleming et al. (2013) in which mentors and mentees were paired, we did not have the information about the paired mentor-mentee relations. However, we determined that this would not be problematic, since Fleming et al. (2013) did not use much of this information in their quantitative analysis.

2.3.2 Exploratory Factor Analysis (EFA)

We assessed the faculty's self-rating and performed exploratory factor analysis (EFA), the similar approach used in the Fleming et al. (2013), to determine the quality of items for measuring research mentoring competency of the BUILD faculty. The instrument developers suggested that the 26 items could be decomposed into 6 highly correlated domains (Table 4 in Fleming et al., 2013). This information suggested that we could try to fit a unidimensional 1-factor model, a 6-factor model or a bi-factor model with one general domain and 6 sub-domains to our survey data.

In the MCA scale, since all the items were coded toward the same direction with a smaller number representing lower frequency or rating and the larger number representing higher frequency or rating, there was no need for recording. Before we ran the EFA, we checked the correlations among items. The responses for psychological items were likert-type ordinal responses, so we chose to use polychoric correlations. After obtaining polychoric correlations, we consulted multiple simple statistical rules of thumb, such as scree plots (Cattell, 1966) and parallel analysis (Horn, 1965) to get a sense about the scale dimensionality. This procedure could help us identify the suitable factorial model and fit a structure that could be easily interpreted (such as Thurstone's Simple Structure [1947], independent cluster, or a bi-factor model). With certain information about the dimensionality, if the information supported the factorial structure reported by the original MCA developer, we would fit three models to the faculty data: a unidimensional 1-factor model, a 6-factor model, and a bi-factor model with one general domain and 6 sub-domains. We then chose the structure based on interpretability and model fit.

2.3.3 Confirmatory Factor Analysis (CFA)

After getting the results from the exploratory analyses on the faculty data, we conducted confirmatory factor analysis on the students' data to check whether the determined dimensionality appeared to be similar. In this process, we assumed that both students and faculty rated the faculty research mentoring competency latent trait rationally. Although students' rated their primary mentors, who might be faculty respondents in our surveys or other faculty in their universities, we observed that 55.3% of the faculty respondents and 57.4% of the student respondents found their mentees or mentors through training programs or departmental assignments. Both students and faculty answered the same questions on the same rating scale at similar periods of time. We believed that this level of matching would provide sufficient evidence that students and faculty were more or less rating similar traits on a similar population.

We fitted the confirmed factorial structure from the previous EFA to the student data, reassessed the dimensionality, and conducted item reduction to create a short form. Quinn (2014) proposed to use the explained common variance ($ECV = \frac{\Sigma\lambda_{ig}^2}{\Sigma\lambda_{ig}^2+\Sigma\lambda_{is}^2}$, where λ_{ig} is the factor loading on the general factor and λ_{is} is the factor loading on a sub-domain) thresholds to specify if the scale is unidimensional or not. With an ECV above .9, the scale could be treated as unidimensional; with an ECV between .7 and .9, we might need extra information to determine the unidimensionality. In practice, Hansen et al. (2014) used item explained common variance ($I\text{-}ECV_i = \frac{\lambda_{ig}^2}{\lambda_{ig}^2+\lambda_{is}^2}$) to assess if the item was strongly associated with the general dimension, and defined that with a general factor loading above .5, the item is strongly related to the general dimension; with an I-ECV above .8, the item is weakly influenced by a group-specific dimension. These criteria were considered for determining the measurement dimensionality and item selection for the short scale.

2.3.4 Evaluation of Short Form Performance

We planned to select a group of items that could well-represent the features (i.e., dimensionality and reliability) of the MCA scale. After item selection, we used multiple procedures, to assess the dimensionality and reliability of the short form, compared to the original MCA scale. We reported the ECV differences to assess the dimensionality, and used Cronbach's Alpha and Squared Multiple Correlation (SMC) to evaluate the reliability of the short form. Additionally, we used Fisher Information to determine how much information an item could provide at each point or how reliable an item could be, and to what extent the item could help with the classification of individuals' latent scores. To get the Fisher Information, we fitted graded response IRT models to the data, and obtained individual IRT scores. Then we checked Pearson's correlation and Spearman's rank order correlation between the scores from the MCA and MCA-short-C. The goal of item selection for creating a short form was to group the most informative items for both faculty and student population that could reflect the same item structure and response structure, and meanwhile, could measure the construct adequately well.

2.4 Results

2.4.1 Exploratory Factor Analysis on Faculty Data

We started with exploratory analyses, using the FAFS data, to examine the potential factorial structures. We firstly produced the polychoric correlation matrix (Table 2.4) from the faculty responses of the 26 items in the MCA. The correlations in the matrix showed that all items were positively and highly correlated (all greater than .37), which somewhat suggested the unidimensional factorial structure. From the item correlation clusters on the diagonal of Table 2.4, We observed high correlations among items that were designed to be in the same sub-domain; meanwhile, some items across different sub-domains were also highly correlated. We ran parallel analyses (Figure 2.1), using this polychoric correlation matrix and its reduced matrix (with SMCs placed on the diagonal), and the results supported the unidimensionality as well.

With the above information, we fitted a unidimensional 1-factor model and chose maximum likelihood (ML) as the factoring method. The results of the 1-factor model (Table 2.5) suggested that all items had relatively high standardized factor loadings (loadings or factor loadings, hereinafter) on one factor (all greater than .66), and the unique variances (uniqueness) of items ranged from .28 to .57. The high factor loadings and communalities (shared variances) suggested that the 1-factor model could be sufficient for measuring one general latent trait, the faculty-student research mentoring competence, and we could roughly conclude that the 26-item MCA scale could be utilized as a reliable measure for the self-assessment of faculty-student research mentoring competence in college settings.

According to Fleming et al. (2013), they designed the MCA to measure 6 aspects of the faculty research mentoring competence, and their confirmed factorial structure was a 6-factor model with highly correlated factors. In our study, we planned to test if the

Table 2.4: Polychoric Correlations

	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	I11	I12	I13	I14	I15	I16	I17	I18	I19	I20	I21	I22	I23	I24	I25	I26
I1																										
I2	.72																									
I3	.72	.70																								
I4	.59	.64	.66																							
I5	.65	.65	.65	.83																						
I6	.45	.52	.45	.53	.58																					
I7	.52	.58	.52	.55	.63	.67																				
I8	.54	.64	.56	.65	.66	.65	.82																			
I9	.57	.55	.56	.62	.63	.51	.58	.71																		
I10	.46	.62	.49	.51	.54	.43	.56	.65	.58																	
I11	.60	.61	.53	.61	.65	.56	.67	.69	.63	.65																
I12	.38	.54	.43	.47	.48	.46	.48	.58	.45	.64	.60															
I13	.45	.58	.45	.47	.49	.45	.52	.58	.47	.69	.62	.87														
I14	.55	.63	.55	.64	.71	.54	.59	.65	.66	.69	.75	.66	.65													
I15	.47	.53	.52	.57	.62	.48	.57	.60	.62	.57	.66	.47	.51	.78												
I16	.56	.57	.61	.61	.61	.46	.53	.64	.58	.56	.71	.51	.51	.71	.71											
I17	.55	.61	.56	.57	.58	.52	.57	.62	.58	.55	.63	.51	.52	.68	.67	.74										
I18	.49	.62	.52	.47	.52	.45	.46	.56	.56	.66	.60	.50	.52	.63	.61	.67	.63									
I19	.52	.61	.57	.60	.62	.56	.59	.66	.55	.62	.70	.60	.62	.70	.66	.70	.69	.71								
I20	.60	.54	.59	.64	.68	.50	.55	.59	.72	.51	.65	.46	.47	.65	.65	.62	.64	.53	.61							
I21	.59	.53	.56	.59	.58	.41	.46	.49	.63	.45	.59	.39	.44	.57	.56	.61	.57	.51	.51	.73						
I22	.43	.45	.55	.50	.54	.47	.52	.53	.50	.39	.53	.37	.37	.55	.56	.59	.58	.50	.63	.53	.58					
I23	.55	.59	.59	.56	.60	.49	.58	.61	.62	.61	.70	.51	.54	.68	.69	.67	.66	.70	.74	.64	.64	.69				
I24	.43	.44	.47	.56	.56	.47	.43	.51	.53	.43	.55	.39	.43	.55	.59	.60	.55	.46	.56	.61	.60	.57	.65			
I25	.54	.55	.62	.63	.69	.59	.57	.64	.64	.53	.62	.44	.47	.65	.68	.71	.68	.60	.68	.71	.64	.66	.74	.69		
I26	.47	.45	.48	.46	.50	.42	.47	.44	.44	.50	.53	.38	.38	.54	.55	.54	.58	.51	.58	.50	.46	.62	.65	.57	.64	

Table 2.5: EFA 1-factor Model (FAFS Sample)

Sub-domains	**Items**	Factor Loading	Communality	Uniqueness
Communication	Listening	0.71	0.50	0.50
	Feedback	0.76	0.58	0.42
	Trust	0.73	0.54	0.46
	Styles	0.77	0.59	0.41
	Strategies (C)	0.81	0.65	0.35
	Coordinate	0.66	0.44	0.56
Expectation	Set expectations	0.74	0.54	0.46
	Align expectations	0.81	0.65	0.35
	Differences	0.77	0.59	0.41
	Goals	0.73	0.54	0.47
	Strategies (E)	0.83	0.69	0.32
Assessing	Knowledge	0.66	0.43	0.57
	Mentee ability	0.69	0.47	0.53
	Mentee skills	0.85	0.72	0.28
Independence	Motivation	0.79	0.62	0.38
	Confidence	0.81	0.66	0.34
	Creativity	0.80	0.63	0.37
	Contributions	0.74	0.55	0.45
	Negotiating	0.83	0.68	0.32
Diversity	Prejudice	0.79	0.62	0.38
	Background	0.72	0.52	0.48
Profession	Network	0.69	0.48	0.52
	Career goals	0.83	0.69	0.31
	Work/life balance	0.70	0.48	0.52
	Role model	0.83	0.68	0.32
	Acquire resources	0.67	0.44	0.56

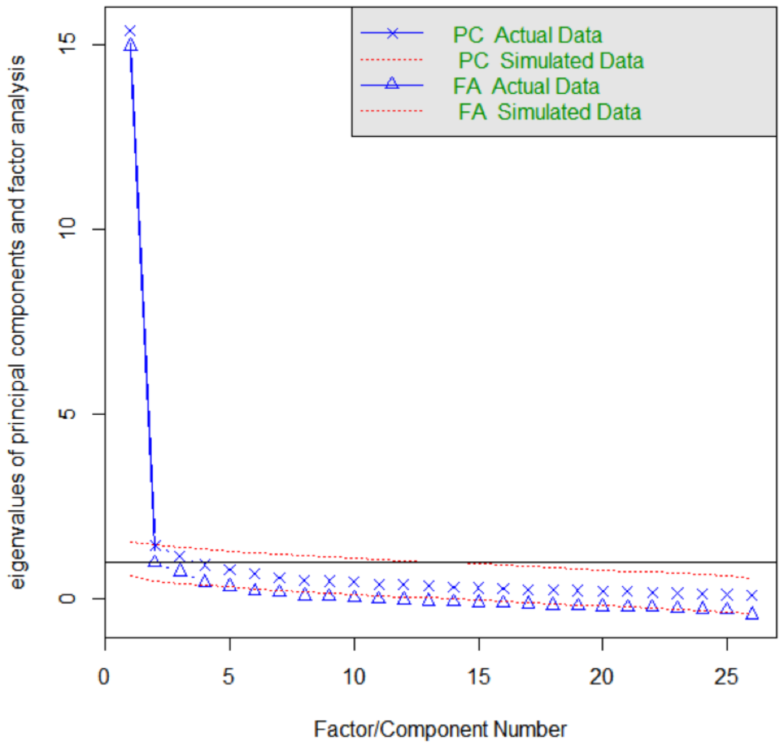


Figure 2.1: Parallel Analysis

factorial structure of 6 competency aspects, or the 6 sub-domains also appeared in our studied populations. Thus, the second EFA model we tried was the 6-factor model that allowed correlations between the factor pairs. In the 6-factor model (Table 2.6), similar to what Fleming et al. (2013) described in their paper, the six sub-domains somewhat stood out. Only four items – "Coordinate," "Goals," "Creativity" and "Negotiating" – did not have factor loadings that were above .3 in their designed sub-domains. Most items were highly loaded on their designed sub-domains. Table 2.6 also presented the relatively high correlations between the factor pairs, and the results were similar to those reported by Fleming et al. (2013). The high correlations between the factor pairs suggested that we could try to fit a bi-factor model.

We explored a bi-factor model, where all 26 items fell under one general mentoring domain and also belonged to a specific sub-domain. By design, we assumed that there were 6 sub-domains, and therefore, we purposefully used a target rotation to lead

Table 2.6: EFA 6-factor Model and Factor Correlations (FAFS Sample)

Item	Communication	Expectation	Assessing	Independent	Diversity	Profession
Listening	0.78					
Feedback	0.72					
Trust	0.73					
Styles	0.40	0.16			0.33	
Strategies (C)	0.37	0.25			0.31	
Coordinate		0.63				0.21
Set expectations		0.85				
Align expectations		0.79				
Differences		0.30			0.47	
Goals			0.40	0.32		
Strategies (E)		0.28	0.20	0.23		
Knowledge			0.92			
Mentee ability			0.93			
Mentee skills			0.31	0.27	0.28	
Motivation				0.32	0.29	0.26
Confidence	0.18			0.35	0.17	0.29
Creativity				0.28		0.30
Contributions	0.19			0.49		0.17
Negotiating		0.19	0.20	0.26		0.37
Prejudice					0.56	
Background	0.22				0.46	0.25
Network						0.68
Career goals				0.27		0.50
Work/life balance					0.28	0.55
Role model		0.17			0.19	0.54
Acquire resources						0.62
Communication	1.00					
Expectation	0.59	1.00				
Assessing	0.58	0.66	1.00			
Independent	0.47	0.56	0.62	1.00		
Diversity	0.56	0.55	0.50	0.38	1.00	
Profession	0.41	0.40	0.40	0.49	0.27	1.00

Note. Only reported factor loadings with absolute values greater than .15

the bi-factor model to fit the structure that all items belonged to a general factor and their designed sub-domains that were proposed by Fleming et al. (2013). All factors in this model were orthogonal to each other. The bi-factor model with six sub-domains was presented in Table 2.7, where we observed a strong pattern that indicated the bi-factor structure, since for all items, their factor loadings on the general domain were greater than .55 and most items loaded majorly on their designed sub-domains.

Among the three EFA models, the bi-factor model was easier to interpret, because the MCA scale was designed to measure faculty mentoring competency and the competency covered 6 aspects. We further compared goodness-of-fit indexes of the 1-factor model, 6-factor model and bi-factor model. We noticed that, for example, the Tucker Lewis Indexes (TLI) increased from .81 in the 1-factor model to .94 in the 6-factor model and then to .96 in the bi-factor model; the root mean squared error of approximation dropped from .11 in the 1-factor model to .06 in the 6-factor model and then to .05 in the bi-factor model. These comparisons along with several other comparisons of indexes indicated that the bi-factor model fitted better than the two other models. Therefore, we chose the bi-factor model as the confirmed factorial structure in the next step.

2.4.2 Confirmatory Factor Analysis on Student Data

We fitted the previously identified factorial structure, the bi-factor model to student data, with the restrictions that items loaded on the general factor as well as a sub-domain proposed by Fleming et al. (2013). Table 2.8 presented the CFA bi-factor model with 6 sub-domains using the SAFS sample. We observed similar features in the CFA results to the EFA target rotation results. All items had high loadings on the general factor (above .71) and weak loadings on a sub-domain (due to the factorial structure constraints). The items' unique variances ranged from .02 to .49, and averaged at .26; the I-ECV ranged from .52 to 1, and averaged at .85. These values indicated that on average around three quarters of the item variances were explained in this model, and the general factor explained large proportions of the explained common variances for each

Table 2.7: EFA Bi-factor Model with 6 Sub-domains under Target Rotation (FAFS Sample)

Item	General	Communication	Expectation	Assessing	Independent	Diversity	Profession
Listening	0.55	0.57			0.16	0.24	
Feedback	0.60	0.52	0.19	0.18	0.23		
Trust	0.60	0.52					0.16
Styles	0.80	0.35					
Strategies (C)	0.84	0.34					
Coordinate	0.63		0.32				0.16
Set expectations	0.66	0.19	0.56				
Align expectations	0.75		0.50				
Differences	0.74		0.26			0.28	
Goals	0.64		0.24	0.31	0.29		
Strategies (E)	0.76		0.24		0.18		
Knowledge	0.63			0.63			
Mentee ability	0.63		0.16	0.70			
Mentee skills	0.84				0.24		
Motivation	0.78				0.27		
Confidence	0.76				0.34		
Creativity	0.71				0.28		0.19
Contributions	0.61				0.49		
Negotiating	0.74				0.29		0.22
Prejudice	0.78					0.35	
Background	0.67					0.47	0.21
Network	0.62						0.48
Career goals	0.72				0.30		0.37
Work/life balance	0.70						0.33
Role model	0.79						0.34
Acquire resources	0.58				0.21		0.42

Note. Only reported factor loadings with absolute values greater than .15

Table 2.8: CFA Bi-factor Model with 6 Sub-domains (SAFS Sample)

Item	General	Communication	Expectation	Assessing	Independent	Diversity	Profession	Uniqueness	I-ECV
Listening	0.71	0.43						0.30	0.73
Feedback	0.74	0.43						0.27	0.75
Trust	0.78	0.42						0.22	0.77
Styles	0.81	0.35						0.22	0.84
Strategies (C)	0.80	0.28						0.28	0.89
Coordinate	0.71	0.05						0.49	1.00
Set expectations	0.79		0.32					0.27	0.86
Align expectations	0.82		0.34					0.21	0.85
Differences	0.81		0.27					0.27	0.90
Goals	0.72		0.03					0.48	1.00
Strategies (E)	0.83		0.13					0.29	0.98
Knowledge	0.74			0.49				0.22	0.69
Mentee ability	0.72			0.69				0.02	0.52
Mentee skills	0.72			0.35				0.35	0.81
Motivation	0.82				0.38			0.18	0.82
Confidence	0.83				0.46			0.09	0.76
Creativity	0.84				0.28			0.22	0.90
Contributions	0.83				0.20			0.27	0.95
Negotiating	0.84				0.10			0.28	0.99
Prejudice	0.84					0.18		0.26	0.96
Background	0.79					0.12		0.36	0.98
Network	0.78						0.34	0.27	0.84
Career goals	0.80						0.40	0.20	0.80
Work/life balance	0.77						0.32	0.31	0.85
Role model	0.83						0.30	0.22	0.88
Acquire resources	0.79						0.31	0.27	0.87

item. The total ECV of the scale performed under the bi-factor model on the student sample was .84, which provided evidence of a strong general factor. For this model, the TLI was .94, the RMSEA was .06, and the comparative fit index (CFI) was .94. These indexes indicated relatively adequate model fit. From the evidence presented above, we could conclude that the bi-factor structure fitted both faculty and student data.

2.4.3 Development and Evaluation of the MCA-short-C

As we noticed in the previous polychoric correlations (Table 2.4), EFA results (Tables 2.5, Table 2.6 and Table 2.7) and CFA results (Table 2.8), the MCA scale was in general unidimensional. Accordingly, we conducted item selections to create a short form that not only inherited features from the original MCA scale performance on both student and faculty population, but could also measure the faculty-student research mentoring competency adequately well. We also acknowledged that the sub-domains proposed by Fleming et al. (2013) mostly appeared in the factorial structure of our confirmed model. We hoped that in the short form, the sub-domains could be well-represented as well. Taking sub-constructs into consideration, we selected items that were well-explained by the general domain and meanwhile, explained by a specific sub-domain by design in the original study by Fleming et al. (2013).

We adapted the selection criteria and procedure from Hansen et al. (2014), with the consideration of prior knowledge about the dimensionality. We planned to retain items that were strongly related to the general factor and weakly influenced by only one of the 6 sub-domains to which they were designed to be associated. The loadings on the general factor (λ_{ig}) and the I-ECV reflected the influence of the general factor. The loadings on the sub-domains (λ_{is}) indicated if the items were also weakly explained by the designed sub-domains. To reflect all 6 sub-domains, we should include at least one item per sub-domains. These criteria should stand for both student and faculty responses, and as a result, we would not select an item if it performed well for one population but not for the other.

In practice, we firstly selected items that had high factor loadings on the general factor in both the EFA bi-factor model and the CFA model, and then we selected those items with high I-ECV. In this process, we excluded items that were not majorly loaded on, in addition to the general factor, their designed sub-domains (e.g., items "Coordinate" and "Goals"). Based on these criteria, we selected one or two most suitable items per sub-domain, and grouped the selected items to create the short form of MCA for measuring college faculty-student research mentoring competency, the MCA-short-C (Tables 2.9).

Table 2.9: Items in the MCA-short-C

Sub-domains	Item	Student			Faculty		
		λ_{ig}	λ_{is}	I-ECV	λ_{ig}	λ_{is}^{*}	I-ECV*
Communication	Styles	0.81	0.35	0.84	0.80	0.35	0.84
	Strategies (C)	0.80	0.28	0.89	0.84	0.34	0.86
Expectation	Align expectations	0.82	0.34	0.85	0.75	0.50	0.69
Assessing	Knowledge	0.74	0.49	0.69	0.63	0.63	0.50
Independent	Motivation	0.82	0.38	0.82	0.78	0.27	0.89
	Confidence	0.83	0.46	0.76	0.76	0.34	0.83
Diversity	Prejudice	0.84	0.18	0.96	0.78	0.35	0.83
Profession	Work/life balance	0.77	0.32	0.85	0.70	0.33	0.82
	Role model	0.83	0.30	0.88	0.79	0.34	0.84

*Only factor loadings on the designed sub-domains were presented and included in I-ECV calculations.

In Table 2.9, we presented the selected items and their λ_{ig}, λ_{is} and I-ECV values from the CFA model on the student sample and the EFA model on the faculty sample. In the selected items, their λ_{ig} ranged from .74 to .84 in the CFA model, and from .63 to .84 in the EFA model; the I-ECVs ranged from .69 to .96 in the CFA model, and from .50 to .89 in the EFA model. Two important points should be mentioned in the selected items. One was that the indexes in the two models were not exactly comparable, even though the results were all standardized, because the EFA model was freely estimated whereas we set constraints in the CFA model. The λ_{is} and I-ECV values from the EFA only considered factor loadings on the designed sub-domains for the selected items. The other important point was that if we did not insist on including at least one item per sub-domain (strictly by design), it would be possible to select items that had higher λ_{ig} and I-ECV values, so that the short form could be even "more" unidimensional. For example, under the "Assessing" sub-domain, the best item out of the three items was "Knowledge" (as we selected into the short form), but this selected item did not perform as well as items selected from other sub-domains. For another example, item "Coordinate" seemed to be a good candidate, if we allowed it to be affiliated with another sub-domain instead of the originally proposed "Communication" domain.

After we obtained the short form, we evaluated the scale's psychometric properties as a whole and compared them to the properties of the original MCA scale (Table 2.10). We compared the ECV of the two scales across the two samples and found that the ECVs of the MCA-short-C were higher than the ECVs of the MCA scale for both student (increased from .84 to .86) and faculty (increased from .74 to .79) samples. The ECVs indicated strong evidence of unidimensionality (Quinn, 2014). Then, we treated both scales as unidimensional and compared common Classical Test Theory (CTT) reliability indexes, the Cronbach's alpha and the SMC (Guttman's lambda 6) values. Although we observed smaller alpha values (.96 to .92 in faculty sample and .97 to .94 in student sample) and SMC values (.97 to .92 in faculty sample and .98 to .94 in student sample) of those obtained from the MCA-short-C scale than those from the MCA scale, the alpha and SMC values of the MCA-short-C still indicated strong reliability, from a classical test theory perspective. We also examined the changes of these two indexes if we dropped any item in the MCA-short-C, and found that the two values would drop at least .02 when excluding any item.

In the Enhance Diversity Study, before responses to the scales like MCA were put into use for evaluating program effects, we assigned scores (usually expected-a-posterior

Table 2.10: Evaluation of MCA-short-C Performance

Comparisons		Faculty	Student
Dimensionality	ECV of MCA	0.74*	0.84
	ECV of MCA-short-C	0.79*	0.86
CTT Reliability .	Cronbach's Alpha of MCA	0.96	0.97
	Cronbach's Alpha of MCA-short-C	0.92	0.94
	SMC of MCA	0.97	0.98
	SMC of MCA-short-C	0.92	0.94
EAP Scores	Pearson's Correlation	0.98	0.98
	Spearman's Rank Order Correlation	0.97	0.98
IRT Reliability**	Empirical reliability of MCA	0.96	0.93
	Empirical reliability of MCA-short-C	0.93	0.89
	Marginal reliability of MCA	0.97	0.92
	Marginal reliability of MCA-short-C	0.93	0.87

*Only factor loadings on the designed sub-domains were included in the ECV calculation.

**Overall reliability.

[EAP] item response scores) to respondents as representations of their abilities of the measured construct. Thanks to this step, consortium-wide researchers could use the scores directly in their data analyses. In addition to the internal consistency of the MCA-short-C, we also compared the scoring alignment between the two scales. Normally, we would use the first survey cohort data or a national sample to obtain the item parameters, and then use the parameters to compute the EAP scores to ensure that the scores could be comparable across responses from different years. In this study, the MCA scale was first-time administered to both populations, and as a result, we had to use the same samples to estimate item parameters and EAP scores on both scales.

We fit two unidimensional graded response IRT models to faculty data and student data separately; one model included 9 items in the MCA-short-C and the other included all 26 items in the MCA. After we obtained item parameters and EAP scores, we used Pearson's correlation and Spearman's rank order correlation to compare the alignment of the scores (Table 2.10). For the faculty sample, Pearson's correlation and Spearman's rank order correlation of the scores from the MCA and the MCA-short-C were .98 and .97, respectively. For the student sample, the correlations were .98 and .98, respectively. These values indicated that the scores from the two scales were highly correlated.

We then compared the overall precision of measurement under the IRT framework, i.e., the overall reliability (ρ_{xx}). The empirical reliability was computed from $\hat{\rho}_{\text{xx}} = \frac{\sigma^2_{S(\hat{\theta})}}{\sigma^2_{S(\hat{\theta})} + \sigma^2_{E(\hat{\theta})}}$, where $\sigma^2_{S(\hat{\theta})}$ was the variance of the estimated scores, and $\sigma^2_{E(\hat{\theta})}$ was computed from the mean squared standard errors of the estimated scores. The marginal reliability was computed from the integration of conditional Fisher information on a given probability density function of the latent trait distribution. In Table 2.10, we reported the comparisons of the overall empirical reliability and marginal reliability of the two scales. For the faculty sample, the empirical reliability dropped from .96 to .93 and the marginal reliability dropped from .97 to .93 when using the MCA-short-C instead of the MCA. Similarly, for the student sample, the empirical reliability dropped from .93 to .89 and the marginal reliability dropped from .92 to .87. The results were congruent with the findings from CTT reliability comparisons. Although the reliability index values decreased when using the short form, the changes were subtle, especially considering the reduced response burden.

To illustrate the relations between the individual scores (standardized to mean 0, variance 1) from the two scales, we plotted the correlations in Figure 2.2a and Figure 2.3a. In these two plots, the diagonal red line represented that the scores obtained from the two scales were equal. The individual scores were all around the red lines for the two samples, though in the student sample, the scores seemed to hit a ceiling effect. We further compared the score distributions of the two scales and plotted the histograms in Figure 2.2b and Figure 2.3b. The corresponding score distributions for the two samples

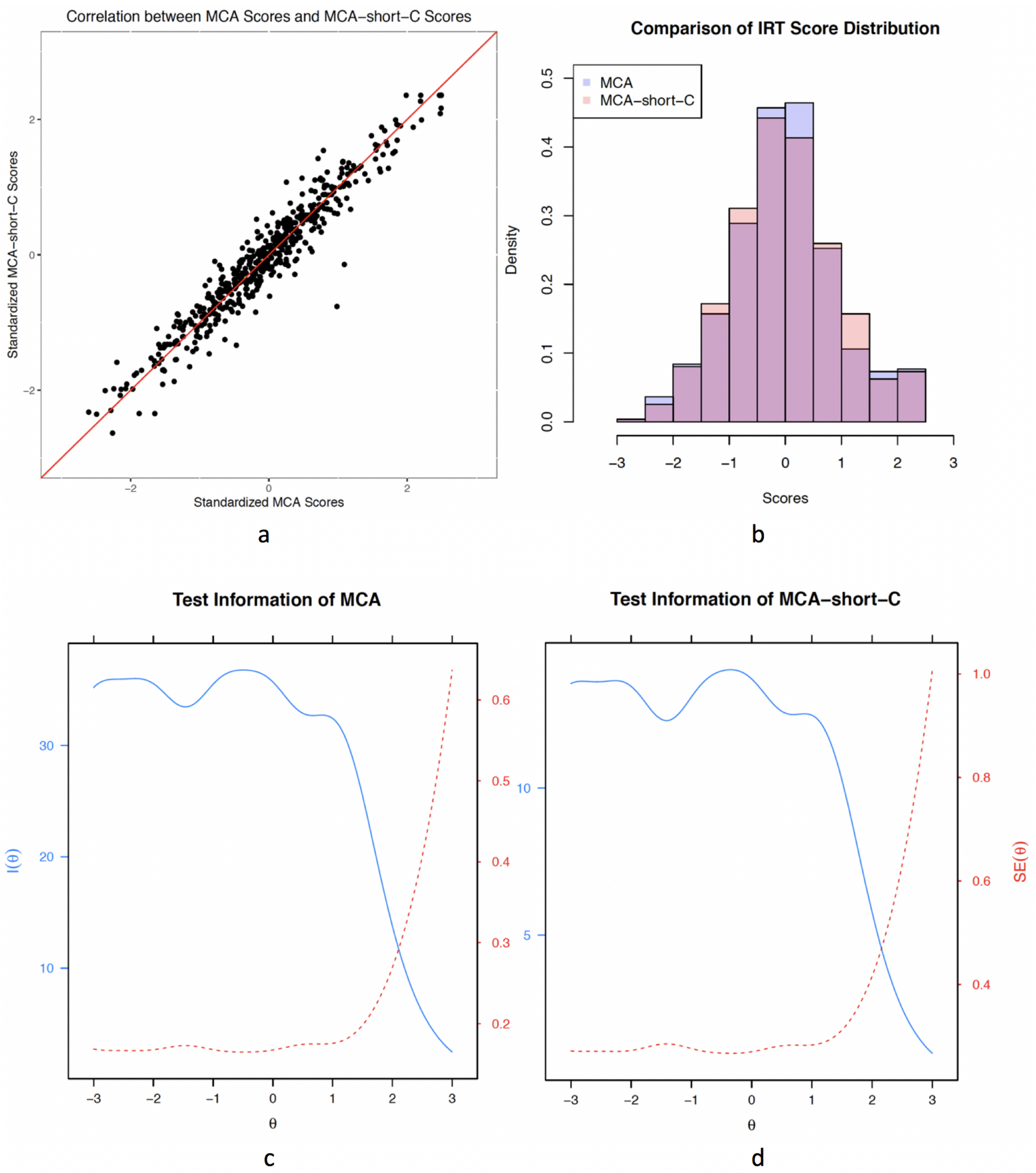


Figure 2.2: Comparisons of MCA and MCA-short-C (Faculty Sample)

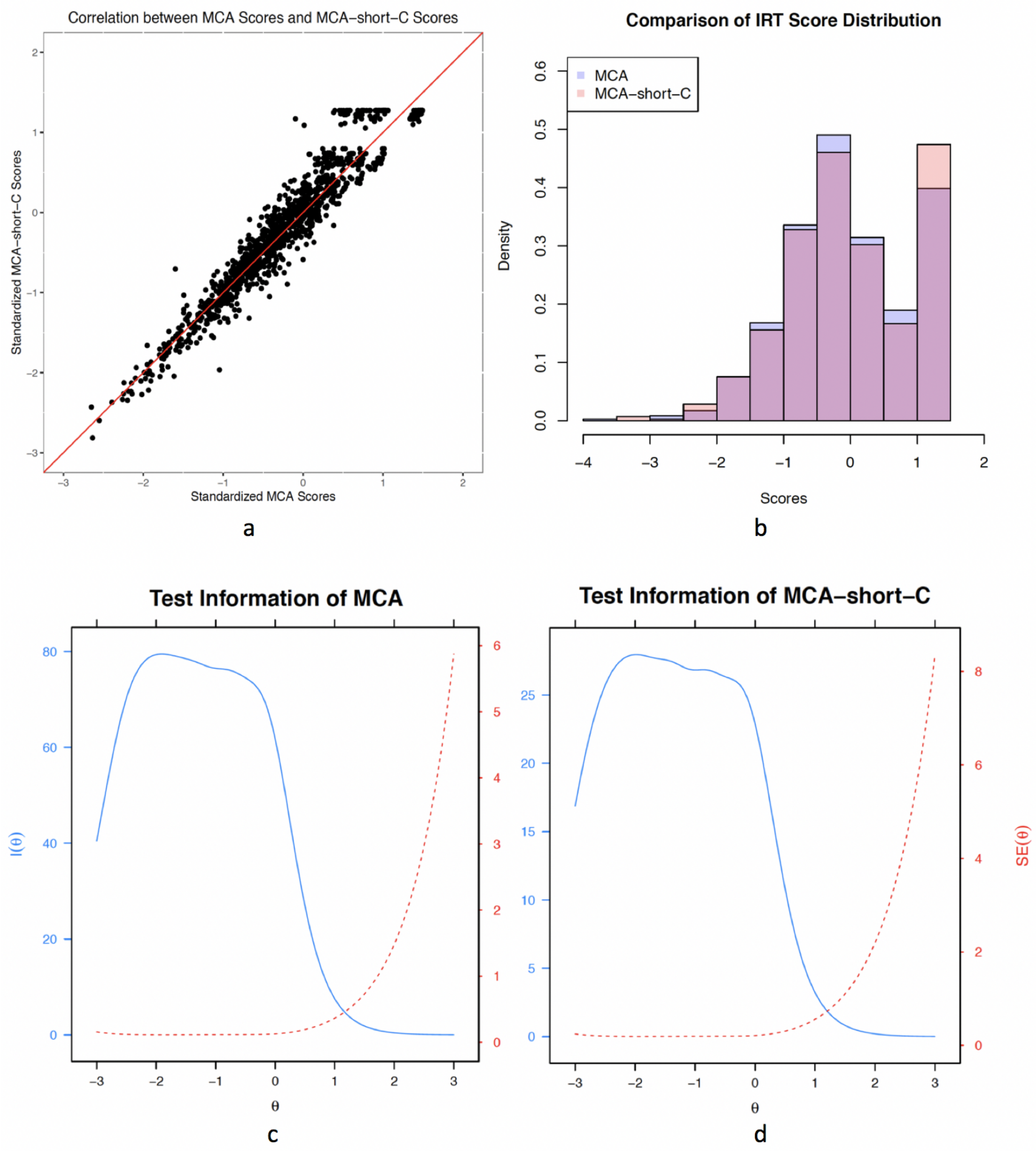


Figure 2.3: Comparisons of MCA and MCA-short-C (Student Sample)

were well aligned, though the student score distributions, especially the MCA-short-C score distribution, were not normal.

We also plotted the test information of the two scales in Figure 2.2c and 2.2d as well as Figure 2.3c and 2.3d. The IRT test information – Fisher information of the scale or the variance of the scores $I(\theta)$ calculated from the sum of item Fisher information, presented the degree of measurement precision of the scales at different ability levels (θ). Although we lost some information when using the MCA-short-C, judging from the similarity of shapes of test information between Figure 2.2c and 2.2d, and between Figure 2.3c and 2.3d, we believed that the MCA-short-C provided sufficient information in the interval where the MCA scale measured the construct reliably.

When faculty respondents' ability (Figure 2.2c and 2.2d) of the latent trait (θ) was smaller than 2 standard deviations (SDs) above the mean, the test information $I(\theta)$ was larger than 6, and the measurement error (the standard error of estimation $\text{SE}(\theta) = \frac{1}{\sqrt{I(\theta)}}$) for the faculty sample was smaller than .4. This indicated that within the same range of θ that covered around 97.5% of faculty respondents, the reliability given θ, $\rho_{\text{xx}|\theta}$ was rather large (over .86), since $\rho_{\text{xx}|\theta} \cong \frac{I(\text{x},\theta)}{1+I(\text{x},\theta)}$ (exact when standardized to variance 1) was monotone increasing. The original MCA scale could maintain a similar level of reliability for 99% of faculty respondents. We concluded that apart from those whose research mentoring scores were at the top 2.5%, the MCA-short-C could reliably measure the construct, and the difference between the scoring reliability conditional on θ was empirically indistinct.

In Figure 2.3c and 2.3d, we noticed that due to the ceiling effect, both MCA and MCA-short-C dropped their measurement accuracy when θ was larger than 1 SD above the mean. Test information of the MCA-short-C scale indicated that the reliability given θ was larger than .78 for around 85% respondents ($\theta < 1.037$). The reliability of the MCA scale given θ was larger than .78 for around 90% respondents ($\theta < 1.283$). Although the reliability of MCA-short-C for high performers was not ideal, this was somewhat inherited from the MCA scale.

2.5 Discussion

2.5.1 Summary of the MCA-short-C

In this study, we investigated dimensionality of the MCA scale, and validated the MCA scale performance on the BUILD faculty and student population. Based on the dimensionality and scale performance, we conducted item reduction and created a short form of MCA, the MCA-short-C (Table 2.9). We performed EFA on the BUILD faculty survey sample from 2017-2018 to explore the factorial structure according to the proposed structure by Fleming et al. (2013), the MCA developer. We confirmed that a bi-factor model with one general domain and 6 sub-domains was the best fit for the faculty data; we then fitted the confirmed structure to the undergraduate student survey data from SAFS 2017. Based on the CFA results, we selected items that had high loadings on the general factor and weakly loaded on one specific sub-domain. These items formed the 9-item MCA-short-C scale as tailored to measure college faculty-student research mentoring.

We utilized multiple approaches to assess the dimensionality, reliability, and scoring congruence of the MCA-short-C, compared with the original MCA scale. We found that the MCA-short-C was unidimensional and reliable for both student and faculty populations. The EAP scores from the MCA-short-C were highly correlated with that from the MCA scale. After assessing the MCA-short-C from different aspects and comparing it to the MCA scale, we confirmed that the MCA-short-C maintained the general measurement properties and functionality of the original MCA and could be used in both faculty and student populations as a substitute for the original MCA scale. The MCA-short-C was reliable for both faculty and student samples, especially when the ability level was not at the top of their group. The MCA-short-C could be offered as an alternative to the long form in measuring faculty-student research mentoring competency, particularly when researchers were concerned about space constraints and/or respondent burden.

The MCA-short-C shared 6 items (”Align expectations,” ”Knowledge,” ”Confidence,” ”Prejudice,” ”Work/life balance,” and ”Role model,” see shared items in Table 2.3 and Table 2.9) with the previously developed MCA-short scale (Zhong et al., 2020). When researchers planned to measure mentoring competency of faculty, especially those whose mentees were mainly graduate students, the MCA-short would be a better option. The MCA-short-C provided the opportunity for assessing undergraduate faculty-student research mentoring from both mentors’ self-rating as well as mentees’ rating on their mentors. Moreover, the MCA-short-C could be a good fit for measuring paired undergraduate faculty-student research mentoring relations. The responses from faculty and students could serve as multiple measures and rating agreement between the two populations could be assessed. To support the Enhance Diversity Study, the item parameters obtained from this study could be used for scoring future participants of the BUILD faculty surveys.

2.5.2 Limitations and Future Studies

Although in this study, we proved that the MCA-short-C was valid for both college faculty and students and the scale reduced survey response burdens, we noticed several drawbacks. First of all, we observed a ceiling effect when students rated their mentors on the MCA-short-C scale. Judging from the scoring on both scales, the ceiling effect might be inherited from the original MCA scale. This indicated that we should put effort toward searching for ”harder” items for the student population to measure the undergraduate faculty-student research mentoring in the future.

Moreover, through IRT modeling, we noticed another measurement issue that was more or less related to the ceiling effect – the collapse of item response categories, especially at the lower ability spectrum. The response category collapse issue appeared in both student and faculty models, and in almost all of the 26 items in the MCA. This suggested that in the future, we should re-examine the item response categories and consider reducing the number of response categories.

The third limitation is that due to the unknown mentor-mentee pairing relations, we could not confirm the rating alignment or agreement. If the pairing relationships could be identified in the Enhance Diversity Study or in another study that collected data using MCA-related scales, we could test the mentor-mentee rating agreement and differential item functioning in the future.

Finally, we mentioned that in addition to the 7 ordinal response options, we also provided an extra ″Not Applicable” or “N/A” option so that participants did not have to respond to the items that were not applicable to their experience. We temporarily treated the “N/A” responses as missing data at random. In the next chapter, we will discuss more about the influence of the “N/A” option.

CHAPTER 3

An Item Response Tree Modeling Approach for Assessing "Not Applicable" Responses in the Enhance Diversity Study

3.1 Introduction

In the Enhance Diversity Study, researchers administer large-scale surveys to collect data for evaluating program effectiveness under the Diversity Program Consortium (DPC). Using surveys to collect data and evaluating program effectiveness starts with finding reliable and valid instruments and understanding the item response process. The previous chapter provided an example of examining the reliability and validity of a scale that was developed by another study and used for measuring the DPC population. This chapter explored the response tendency or item response decision-making process of the DPC population. Considering the importance of faculty mentoring in the DPC, we continued to use measuring faculty mentoring competency as an example to explore the item response process. Building upon the study in Chapter 2, we analyzed faculty and student responses to the 26-item Mentoring Competency Assessment (MCA) scale (Fleming et al., 2013), with a purpose of understanding the meaning of "Not Applicable" (or "N/A") which was provided as an additional response category in many survey items in the DPC survey administration.

The DPC surveys provided the "N/A" option as a response category so participants could have more opportunities to express their actual conditions. However, how to interpret "N/A" or other similar response alternatives has rarely been studied. These responses were often treated in analytical models as missing at random, although by design, participants were provided opportunities to distinguish the use of “N/A” from a missing response. In this study, we used an item response tree modeling approach (IRTree, De Boeck & Partchev, 2012; Jeon & De Boeck, 2016) to assess the influence of ”N/A” responses and to explore potential response processes when participants selected their response options.

3.1.1 Response Alternatives

Item response alternatives are the choices or response categories that participants could choose when responding to an item. In a Likert-type scale, the alternatives are usually coded as ordinal responses. In addition to ordinal responses, response alternatives could also be response categories that had no ordinal indication, such as “not applicable,” “don’t know,” or ”prefer not to state.” Although the reasons for choosing this type of responses might be unknown, these options could help reduce forced choosing and non-response (Oldendick, 2012).

Survey method researchers encouraged the use of “not applicable” type of response options (Holman, Glas, Lindeboom, Zwinderman, & De Haan, 2004; “Don’t knows (DKs)”, n.d.; Oldendick, 2008); however, how to analyze ”N/A” responses beyond descriptions has rarely been studied. Berinsky and Margolis (2011) urged that researchers should treat such responses as “not applicable” with caution, and used descriptive data to conclude that participants who were socioeconomically disadvantaged tend to choose this type of options in the polling and health care surveys. Apart from presenting descriptive frequency, “not applicable” and “don’t know” responses were often treated as non-substantive responses, and in data analysis and statistical modeling, as missing data (Holman et al., 2004; “Don’t knows (DKs)”, n.d.; Berinsky & Margolis, 2011). Holman et al. (2004, p. 29) proposed 4 approaches for handling “N/A” responses; however, all 4 approaches, i.e., “cold deck imputation, hot deck imputation, treating the missing responses as if these items had never been offered to those individual patients, and using a model which takes account of the ‘tendency to respond to items’," were merely treating “N/A” as missing responses. Limited by available analytical approaches, our understanding of ″N/A″ type of responses largely stays at the descriptive level. It is possible that the ″N/A″ type of responses are nothing more than missing responses at random or representing a lowest response category in some cases, while it is also possible that the ″N/A″ responses could cause missing representation or even misrepresentation of certain groups of respondents. Analytical approaches should be implemented to help researchers dig more information from ″N/A″ type of responses.

In the measurement field, researchers put great efforts on accounting for the “don't know” condition in answering test and assessment questions (Bock, 1972; Samejima, 1979; Thissen & Steinberg, 1984), when participants did not know the answer yet correctly responded to a multiple choice question. Although the setting was not the same as the survey response conditions, previous research indicated the possibility of using measurement models to analyze ″N/A″ type of responses.

3.1.2 ″Not Applicable” Options in the Enhance Diversity Study

In the Enhance Diversity Study, researchers identified existing scales to measure the Hallmarks. Some scales were not necessarily designed for populations that were similar to the DPC program participants, and as a result, some items might not be applicable for all DPC participants. Researchers noticed this issue, and added “N/A,” ″prefer not to state,″ ″other,″ and ″don't know″ options as additional response categories to provide respondents with opportunities to declare their actual conditions. Researchers might revise items or add additional response alternatives if they noticed large percentages of missing responses in certain survey questions from a previous year. The purpose was to make the items more friendly to respondents, respect their thoughts, avoid forced choosing and collect more accurate information. The “N/A” type of response alternatives leveraged the difficulty level in data analyses, and even though we collected extra information from those responses, we did not understand their actual influence. In this study, we took the added "N/A" response option in the MCA scale as an example, and explored the meaning of "N/A" responses.

The MCA scale (Fleming et al., 2013) was developed for evaluating the research mentoring in the NIH Clinical and Translational Science Awards (CTSA, Pfund et al., 2013, 2014) program. Participants in the CTSA were clinical professionals and researchers, and the MCA was originally developed for the CTSA participants in the clinical training settings. To respect the actual conditions of the DPC participants, when utilizing the MCA scale to measure research mentoring of the DPC program participants, an "N/A" response option was added to each item. In other words, items in the MCA scale administered by the DPC had 8 different response categories; they were numeric responses 1 – 7, representing participants' rating of mentoring skills from "Not at all Skilled" (1) to "Moderately Skilled" (4) and to "Extremely Skilled" (7), plus one "N/A" option, representing the activity that the item described was not applicable to the respondents. The "N/A" response option was a special case, since it did not have ordinal meaning and might lead to unique response patterns.

Although the "N/A" response option was provided to participants and we noticed large percentages of choosing "N/A" in certain items, we often treated the "N/A" responses as missing responses at random (e.g., the study in Chapter 2). Since we provided the "N/A" response option in the surveys, we hoped that we could know more about the meaning of the "N/A" responses and how we could treat the "N/A" responses in the future statistical analyses.

3.2 Item Response Tree Modeling Approach

3.2.1 Item Response Theory Modeling for Analyzing Latent Traits

Abstract and socially constructed concepts (Hacking, 1997), such as faculty research mentoring competency, are latent traits or latent constructs in measurement models. Latent traits, being largely used in social sciences as predictors (Schofield, 2015), are unobserved random variables in a statistical model that cannot be directly measured, but they associate with and can be modeled by observed variables or manifest variables (Bollen, 2002; Cai, 2012). In the field of measurement, a common belief is that the observed behaviors are caused or influenced by underlying latent traits (Bollen, 2002). This belief leads to the approach for measuring latent traits – using a set of items to obtain observable data to assess a latent trait (Thurstone, 1925). Scales like the MCA were developed for measuring latent traits and measurement models, such as the Item Response Theory (IRT) models were widely used for analyzing and scoring items that measured latent traits (Jeon & De Boeck, 2016; Zhang, 2016).

The IRT models are logistic models, and can be formulated to illustrate the probability of selecting a certain response to an item, conditional on participants' ability of a latent trait. IRT models provide item level information through the discrimination parameters and difficulty parameters. The difficulty parameter describes the location where the amount of the latent trait has a .5 probability of endorsing the item. There can be multiple locations if an item has polytomous responses. The discrimination parameter is the slope of the curve at the item location. In an item with ordered polytomous categories, the discrimination parameter is the same across the ordinal responses, under the graded response models. The IRT parameters can be expressed in a form of generalized linear mixed models (GLMM, De Boeck & Partchev, 2012). The GLMM parameterization is more commonly seen in statistical modeling, although the parameters seem to be less intuitive in item plots.

Although traditional IRT models could provide plenty of information about the performance of items and overall scales, the models would not provide information related to the response process or response decision making. For example, when DPC survey participants responded to an item in the MCA, they might firstly consider if the item was applicable to them, and if so, they would then declare their skill levels. This process could not be reflected, if we only fitted a graded response model and treated the ″N/A″ responses as missing at random.

3.2.2 Item Response Tree Models

De Boeck and Partchev (2012) as well as Jeon and De Boeck (2016) proposed to use the item response tree (IRTree) model, an item response model with a response tree structure, to analyze response processes. The IRTree model is defined as "a postulated internal decision process with a tree structure, which is composed of sub-trees and their corresponding nodes and branches" (p. 1070, Jeon & De Boeck, 2016). In the IRTree models, researchers add a node into the tree structure, at the point where they believe a decision making behavior happened when participants chose a response category over another. The IRTree modeling could be a potential method for handling such responses as "not applicable," and was implemented in this study.

Jeon and De Boeck (2016) presented an IRTree structure in their Figure 3 (p. 1077), which modeled item response process of choosing an option on verbal aggression items (Smits, De Boeck, & Vansteelandt, 2004) from a three-point Likert-type scale with the "No," "Perhaps" and "Yes" categories. In this IRTree, the first Node Y_1^* portrayed the decision process that the participants chose not to present verbal aggression, or possibly gave an aggressive verbal response to a particular scenario. For those who chose to cast verbal aggression, their next response step was to make a decision on second Node Y_2^* to indicate the level of certainty. For those who chose "No" on the first Node, their responses on the second Node would be marked as missing information. From a modeling perspective, instead of fitting a graded response model (as we normally would) to the data, Jeon and De Boeck (2016) fitted a multi-dimensional IRT model and estimated the correlation between the nodes Y_1^* and Y_2^*. When they observed a small correlation, they concluded that depending on the chosen response options, different latent variables were measured. The multi-dimensionality held, and the response options were not ordinal (Smits et al., 2004).

3.2.3 Illustration of the IRTree Modeling

To better illustrate the IRTree modeling, we used a possible IRTree structure of responding to the MCA scale as an example to demonstrate the IRTree modeling approach as well as its advantages and potential issues. We also explained the IRTree modeling from a statistical modeling perspective to articulate the relations between IRTree models and traditional IRT models.

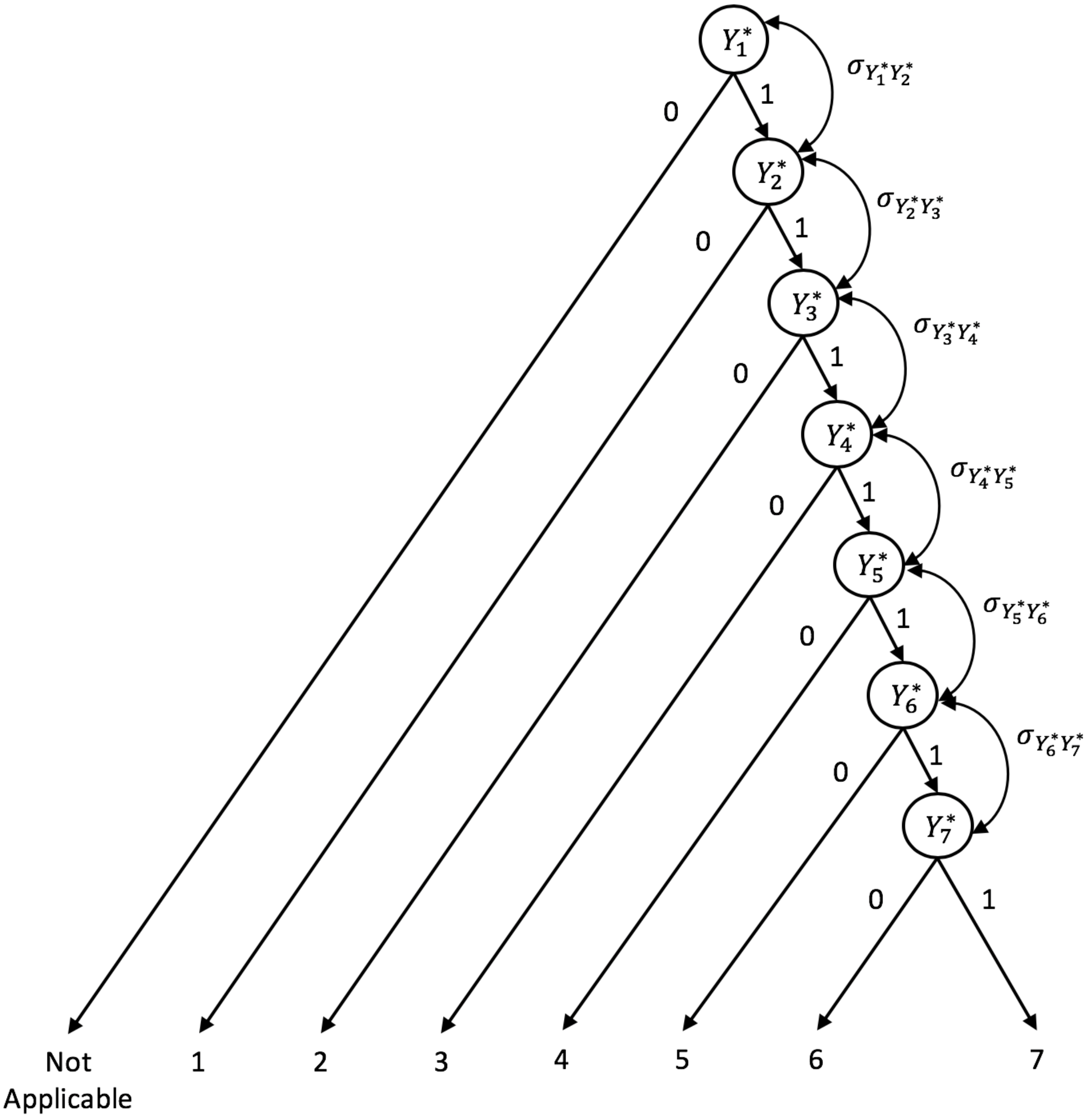


Figure 3.1: Illustration of an IRTree Structure of Responses to the MCA Scale

Treating the MCA scale as unidimensional and regardless of the actual missing responses, the most complicated IRTree structure, presented in Figure 3.1, consists of seven nodes where each node is associated with two branches. Each node represents the decision making process of choosing the left branch (coded as 1) or the right branch (coded as 0). Node 1 (Y_1^*) distinguishes choosing "Not Applicable" or the seven ordinal response categories; Node 2 (Y_2^*) separates choosing category "1" versus choosing other categories above 1; similarly, Node 3 (Y_3^*) separates choosing "2" versus choosing other categories above 2, and so forth, until Note 7 (Y_7^*) that separates choosing the category "6" versus choosing the category "7". This IRTree structure treats the "N/A" response option as the first decision to make, or the lowest response category that respondents would first consider.

In this IRTree structure (Figure 3.1), we assume that all nodes are correlated. To test the necessity of having a particular node in the structure, we need to look at if the values of correlations between node pairs, especially correlations between the directly attached nodes (e.g., $\sigma_{Y_1^*Y_2^*}$, $\sigma_{Y_2^*Y_3^*}$, . . ., $\sigma_{Y_6^*Y_7^*}$). If a node is highly correlated with one of its neighbors, then it would not be necessary to have the two nodes, and the responses could be treated as ordinal. For example, if $\sigma_{Y_2^*Y_3^*}$, $\sigma_{Y_3^*Y_4^*}$ . . ., $\sigma_{Y_6^*Y_7^*}$ are all high, we could treat response options 1 – 7 as ordinal. If a node is barely correlated with one of its neighbors, then the decision making at the two nodes are almost independent. For instance, if $\sigma_{Y_1^*Y_2^*}$ is close to 0, we could treat the "N/A" responses as missing at random.

A mapping matrix can mimic the choices of branches at each node to pseudo-item responses as if the responses were from a regular multidimensional IRT model. Figure 3.2 presents the mapping matrix of the IRTree structure in Figure 3.1. The mapping matrix T in Figure 3.2 shows how "regular" item responses Y_{pi} (where p is "person," and i is "item") in rows are linked to the pseudo-item responses in columns. For the observed outcome $Y_{pi} = N/A$, it corresponds to (Y_{pi1}^*, Y_{pi2}^*, Y_{pi3}^*, Y_{pi4}^*, Y_{pi5}^*, Y_{pi6}^*, Y_{pi7}^*) = (0, NA, NA, NA, NA, NA, NA), where NA represents a missing observation, since $Y_{pi} = N/A$ does not involve responses in other nodes. The interpretation of other observed outcomes are rather similar.

	Y^*_{pi1}	Y^*_{pi2}	Y^*_{pi3}	Y^*_{pi4}	Y^*_{pi5}	Y^*_{pi6}	Y^*_{pi7}
$Y_{pi} = N/A$	0	NA	NA	NA	NA	NA	NA
$Y_{pi} = 1$	1	0	NA	NA	NA	NA	NA
$Y_{pi} = 2$	1	1	0	NA	NA	NA	NA
$Y_{pi} = 3$	1	1	1	0	NA	NA	NA
$Y_{pi} = 4$	1	1	1	1	0	NA	NA
$Y_{pi} = 5$	1	1	1	1	1	0	NA
$Y_{pi} = 6$	1	1	1	1	1	1	0
$Y_{pi} = 7$	1	1	1	1	1	1	1

Figure 3.2: Mapping Matrix

Each node in this IRTree structure, apart from the first one, is conditional on the previous node. According to Jeon and De Boeck (2016), the conditional probability of the (m, k)-th element of the mapping matrix T, $\Pr(Y^*_{pik} = T_{mk}|\theta_{pk})$, where m represents the m-th terminal observed outcome or the m-th row in matrix T and k represents the k-th node or the k-th column in matrix T, can be formulated as (Eq. 3.1):

$$\Pr\left(Y^*_{pik} = T_{mk}|\theta_{pk}\right) = g^{-1}(\alpha_{ik}\theta_{pk} + \beta_{ik}), \tag{3.1}$$

In Eq. 3.1, $g(\cdot)$ is the logit link function, α_{ik} is the item slope and β_{ik} is the intercept for item i node k. The parameters α_{ik} and β_{ik} are item parameters in a 2-parameter logistic model (2PL) under the GLMM framework notation. Then the model for observed terminal outcome $Y_{pi} = m$ can be expressed as (Eq. 3.2):

$$\Pr\left(Y_{pi} = m|\theta_{p1}, \ldots, \theta_{p7}\right) = \prod_{k=1}^{7} \Pr(Y^*_{pik} = T_{mk}|\theta_{p1}, \ldots, \theta_{p7})^{tmk}, \tag{3.2}$$

In Eq. 3.2, $k \in (1, 2, \ldots, 7)$ represents the node, $m \in (1, 2, \ldots, 8)$ represents the response category, θ_{pk} is the latent trait at node k, and $\boldsymbol{\theta}_p = (\theta_{p1}, \ldots, \theta_{p7})'$ follows a multivariate distribution with $\boldsymbol{\theta}_p \sim N(\mathbf{0}, \Sigma)$, where Σ is a 7×7 covariance matrix. The above description indicated that although the logic of modeling differs between the IRTree modeling and traditional IRT modeling, the estimation and parameterization of the two are rather similar.

3.3 Methods

As mentioned earlier, the illustrated structure (Figure 3.1) is the most complicated IRTree structure. The number of nodes reflects the number of dimensions, and estimating high-dimensional models often causes heavy computational burden due to the curse of dimensionality. This reminds us to use prior knowledge to design potential IRTree models that are simple and interpretable. In this study, we used findings from the study in Chapter 2 and re-coded the response categories. Based on the new categories, we proposed potential IRTree structures.

3.3.1 Data and Response Patterns

In this study, similar to the study in Chapter 2, we included faculty/mentor and student/mentee responses to the MCA scale from the BUILD Faculty Annual Follow-up Survey (FAFS, 2017-18) and the BUILD Student Annual Follow-up Survey (SAFS, 2017) as our analytical data. In addition, as a majority of mentors in the DPC programs were in the NRMN programs, we combined the BUILD Faculty Survey (2017-18) data and the NRMN Survey (2018) data together as the faculty/mentor sample. The NRMN faculty survey was distributed to all NRMN participants, and received 1107 responses. Among them, 754 participants self-identify as a mentor, and 621 participants responded to at least one MCA item on the ordinal scale of 1 – 7. The BUILD survey samples were described in Chapter 2. Combining the NRMN faculty mentors and the 565 faculty

Table 3.1: N/A Response Patterns

Item	”N/A” Responses	
	Faculty (1156)	Student (1760)
Listening	0	192
Feedback	0	187
Trust	1	202
Styles	4	215
Strategies (C)	9	237
Coordinate	215	474
Set expectations	25	269
Align expectations	23	238
Differences	23	268
Goals	52	375
Strategies (E)	5	226
Knowledge	40	363
Mentee ability	46	415
Mentee skills	11	456
Motivation	5	231
Confidence	3	232
Creativity	14	272
Contributions	22	320
Negotiating	69	345
Prejudice	23	437
Background	8	338
Network	27	344
Career goals	20	306
Work/life balance	21	356
Role model	8	295
Acquire resources	59	278

mentors who rated at least one MCA item on the ordinal scale of 1 – 7, the total sample of the faculty mentors was 1156. Similarly, the student mentee sample included 1760 participants who responded to at least one MCA item on the ordinal scale of 1 – 7.

In addition to the ″N/A″ responses, we observed missing responses in 31 faculty participants; none of the student participants had missing responses. Since the number of missing responses was small, we treated them as missing at random or ignorable in the analysis. We reported the N/A response pattern in Table 3.1. The ″N/A″ response patterns in the Table 3.1 showed that the proportions of ″N/A″ responses were relatively small in the faculty sample across all items, apart from the ″Coordinate″ item. In some items, such as ″Listening,″ ″Feedback,″ ″Trust,″ and ″Confidence,″ the proportions of ″N/A″ responses were none or close to none. There were more ″N/A″ responses in the student responses, the proportions of ″N/A″ responses across all items were more consistent than those in the faculty responses.

In Chapter 2, we found that the response category collapse issue appeared in both student and faculty populations, and proposed to reduce the number of response categories in future studies. We also noticed that not all ordinal response categories appeared in the responses. For example, the ordinal response ″1″ did not appear in the item ″Feedback.″ In the MCA scale, the numeric response categories 1 – 7 represented participants' rating of mentoring skills from "Not at all Skilled" (1) to ″Moderately Skilled" (4) and to "Extremely Skilled" (7). As our study served the purpose of providing an analytical demonstration, to simplify the analytical models and to avoid missing response categories, we re-coded the the numeric response categories 1 – 7 to "Not at all Skilled" (1 – 2), ″Moderately Skilled" (3 – 5), and "Extremely Skilled" (6 – 7). Analyses in this study were based on the re-coded response categories. We presented the corresponding relations between the original response categories and the re-coded response categories in Table 3.2.

Table 3.2: Re-coded Response Categories

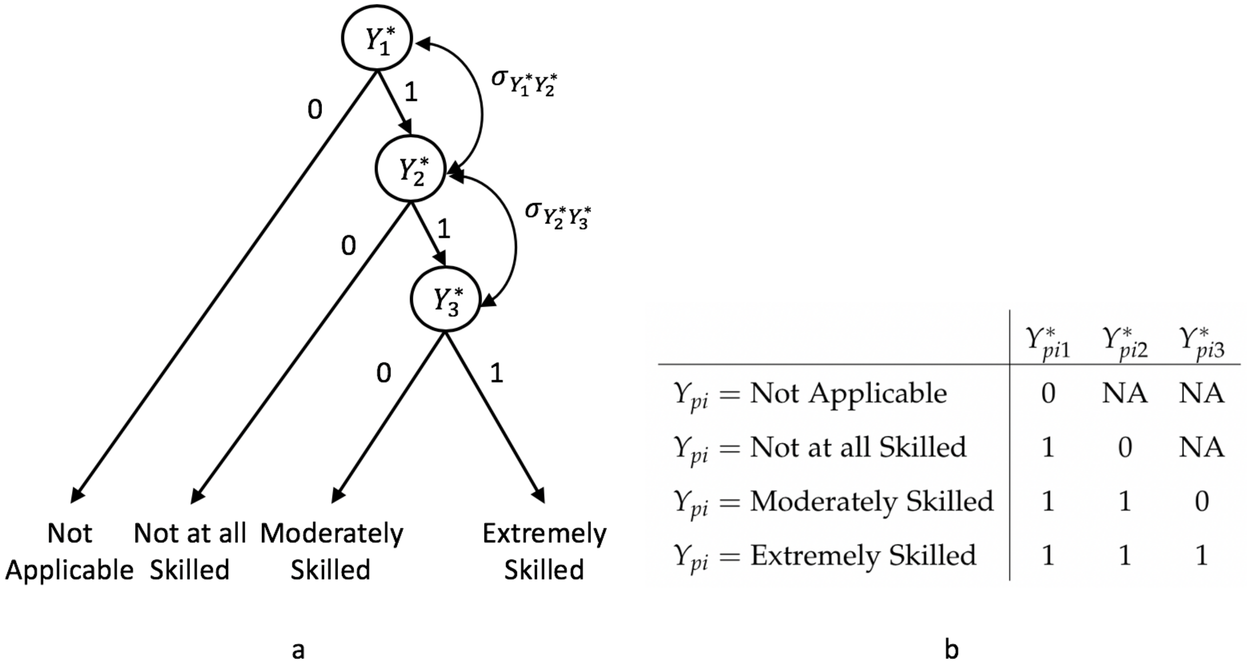
Re-coded Categories	Original Response Categories
Not at all Skilled	1
	2
Moderately Skilled	3
	4
	5
Extremely Skilled	6
	7
Not Applicable	Not Applicable

3.3.2 Proposed IRTree Models

From the findings in Chapter 2 and in Zhong et al. (2020), we knew that the MCA scale could be considered as a unidimensional scale for both faculty and students. Our focus in this study was to use the IRTree modeling to explore the meaning of ”N/A” responses, so we considered the unidimensional measurement structure as the confirmed structure, even after we re-coded the ordinal responses.

3.3.2.1 Model One

The first proposed IRTree model was a 3-node model that was similar to the model we illustrated in Figure 3.1, but simplified due to the reduced number of response categories. In Figure 3.3, we presented the IRTree structure (3.3a) and the mapping matrix (3.3b) of Model 1. In this model, we assumed a three-stage decision process: respondents would decide if this item was applicable to them at the first stage (node Y_1^*); if the item was applicable, participants would decide if the mentor was skilled in this item at the second stage (node Y_2^*); if the mentor was skilled, then participants would decide the skill level – moderately skilled or extremely skilled, at the final stage (Y_3^*). In



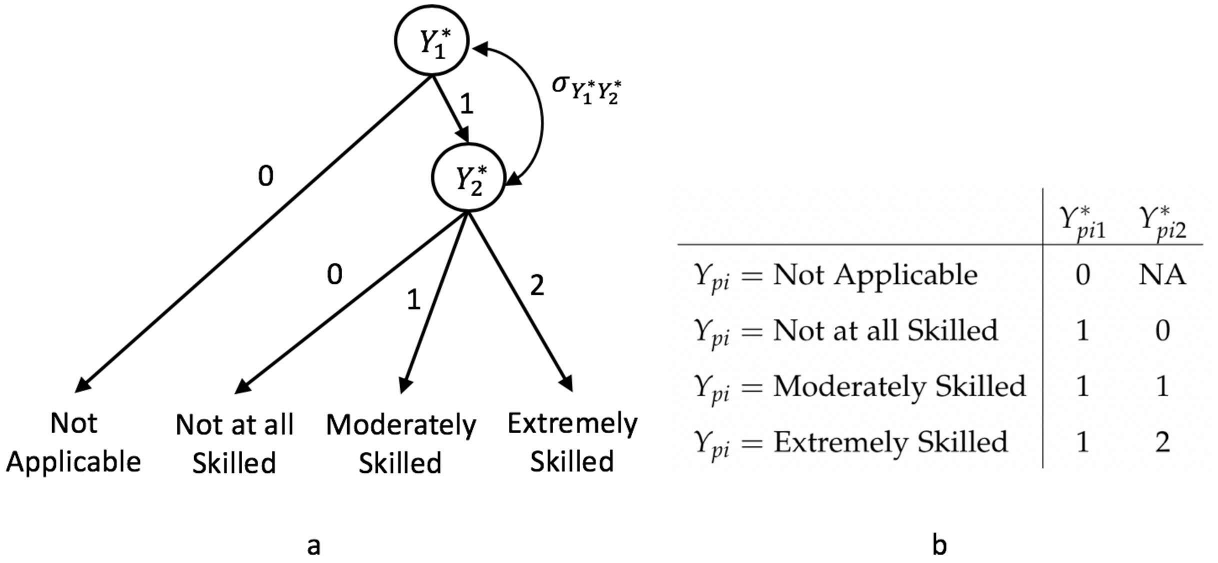
	Y_{pi1}^*	Y_{pi2}^*	Y_{pi3}^*
Y_{pi} = Not Applicable	0	NA	NA
Y_{pi} = Not at all Skilled	1	0	NA
Y_{pi} = Moderately Skilled	1	1	0
Y_{pi} = Extremely Skilled	1	1	1

Figure 3.3: IRTree Structure and Mapping Matrix of Model 1

this process, if participants chose a left branch at Y_1^* or Y_2^*, their choices at later nodes would be automatically coded as missing. The model could be considered as a three dimension 2PL model that allowed the dimensions to be correlated; each dimension included 26 items.

3.3.2.2 Model Two

Model 2 (Figure 3.4) was a 2-node model that mapped a two-stage decision process: respondents would decide if this item was applicable to them at the first stage (Y_1^*); if the item was applicable, then participants would decide the skill level – not at all skilled, moderately skilled or extremely skilled, at stage two (Y_2^*). In Figure 3.4, we presented the IRTree structure (3.4a) and the mapping matrix (3.4b) of Model 2. The model could be considered as a two dimension model in which one dimension included 26 items that parameterized as 2PL, and the other included 26 items that parameterized as graded response model; the two dimensions were correlated.



	Y^*_{pi1}	Y^*_{pi2}
Y_{pi} = Not Applicable	0	NA
Y_{pi} = Not at all Skilled	1	0
Y_{pi} = Moderately Skilled	1	1
Y_{pi} = Extremely Skilled	1	2

Figure 3.4: IRTree Structure and Mapping Matrix of Model 2

3.3.3 Analysis

In addition to the IRTree models, we fit the unidimensional graded response IRT model to both faculty and student data as a reference model. We used the Maximum Likelihood (20 quadrature points) to estimate the proposed tree models, and similar to the estimation of traditional IRT models, the Full-Information Maximum Likelihood (FIML) could help with handling missing data. The missing data included a small portion of missing data from the original dataset, and ″planned″ missing data due to choosing a left branch before the final node. Model estimation and scoring would be based on complete items responses, and missing responses, although unable to contribute extra information, would not be problematic.

In Table 3.1, we noticed that there was no ″N/A″ observation in the item ″Listening″ and ″Feedback″ in the faculty sample. We were unable to estimate the parameters if the response categories were not observed. To make the model estimation possible, we added a ″fake″ response record that selected ″N/A″ in the item ″Listening″ and ″Feedback,″ selected ″Moderately Skilled″ in the item ″Trust″ (or rather, any other items apart from ″Listening″ and ″Feedback″), and responses to all other items were missing. We assumed that this ″fake″ participant who believed that the item ″Listening″ and ″Feedback″ were not applicable, had moderate skills in the item ″Trust,″ and the participant's skills in other items were estimated based on the moderate skills in the item ″Trust″ using FIML. With this ″fake″ response record, we were able to estimate the models, although the item parameters for the ″N/A″ responses in the item ″Listening″ and ″Feedback″ were meaningless and the scoring for this participant was unstable. Interpreting item parameters for the ″N/A″ responses was not a focus in this study, and item scoring for this ″fake″ participant was ignored since this participant did not exist in reality.

After model estimation, we examined correlations between nodes, and evaluated the necessity of including the nodes in the models. We used information based statistics, e.g., the Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) to compare the relative model fit of Model 1, Model 2 and the corresponding unidimensional graded response IRT model.

3.4 Findings and Implications

We fitted the Model 1, a 3-node model to the student and faculty data, and reported the correlations in Table 3.3. We noticed that for both faculty and students, the correlation between Y_2^* and Y_3^* ($\sigma_{Y_2^* Y_3^*}$) was relatively high. If the $\sigma_{Y_2^* Y_3^*}$ was close to 1, the nodes Y_2^* and Y_3^* could be combined, and the skill levels – not at all skilled, moderately skilled, and extremely skilled – were ordinal. However, $\sigma_{Y_2^* Y_3^*}$ was not high enough, say, over .8, and that Y_2^* and Y_3^* should still be considered as measuring two separate dimensions. The correlation between Y_1^* and Y_3^* ($\sigma_{Y_1^* Y_3^*}$) was almost 0, which confirmed our assumption of the order of response decision making.

Table 3.3: Node Correlations (Model 1)

	Faculty			Student		
	Y_1^*	Y_2^*	Y_3^*	Y_1^*	Y_2^*	Y_3^*
Y_1^*		$\sigma_{Y_1^*Y_2^*}$	$\sigma_{Y_1^*Y_3^*}$		$\sigma_{Y_1^*Y_2^*}$	$\sigma_{Y_1^*Y_3^*}$
Y_2^*	-0.285		$\sigma_{Y_2^*Y_3^*}$	0.01		$\sigma_{Y_2^*Y_3^*}$
Y_3^*	-0.007	0.534		0.036	0.621	

The most important correlation, $\sigma_{Y_1^*Y_2^*}$ was slightly different in faculty and student data. In the student data, the $\sigma_{Y_1^*Y_2^*}$ was close to 0, which indicated that the "N/A" responses could be treated as missing at random. In the faculty data, the $\sigma_{Y_1^*Y_2^*}$ was -.285, and although the value was small, we could not ignore the correlation. Thus, "N/A" responses could not be treated as missing at random for the faculty responses. The negative correlation somewhat indicated that choosing "N/A" responses was related to the declaration of having research mentoring skills. In other words, participants who tended to choose the "N/A" option were more likely to choose skilled in the measured mentoring competency.

Table 3.4: Node Correlations (Model 2)

	Faculty		Student	
	Y_1^*	Y_2^*	Y_1^*	Y_2^*
Y_1^*		$\sigma_{Y_1^*Y_2^*}$		$\sigma_{Y_1^*Y_2^*}$
Y_2^*	-0.303		-0.382	

Table 3.4 presented the node correlations in Model 2, which were estimated from fitting the 2-node model to the student and faculty data. In both student and faculty samples, the we observed negative correlation between nodes Y_1^* and Y_2^* ($\sigma_{Y_1^*Y_2^*} = -0.382$ in the student sample, and $\sigma_{Y_1^*Y_2^*} = -0.303$ in the faculty sample). The "N/A" responses were not ignorable and could not be treated as missing at random for the faculty responses. The negative correlation suggested that participants who tended to choose the

”N/A” option were more likely to rate higher skill levels in research mentoring. For the student sample, the estimated $\sigma_{Y_1^*Y_2^*}$ in Model 2 led to a different conclusion than what we concluded from Model 1.

Table 3.5: Model Comparisons

	Faculty		Student	
	AIC	BIC	AIC	BIC
IRTree Model 1	38658.65	39462.17	47509.49	48379.70
IRTree Model 2	40636.97	41298.99	51008.73	51725.70
Unidimensional IRT Model	36005.94	36400.12	35333.04	35759.93

We compared the two IRTree models with the unidimensional IRT model that treated the ”N/A” responses as missing data at random. The unidimensional IRT model was nested within the Model 2, and both the unidimensional IRT model and Model 2 were nested within Model 1. Judging from AIC and BIC values (Table 3.5), the unidimensional IRT model fitted the best, and the Model 2 fitted the worst. Model 1 fitted better than Model 2, which again, indicated that Y_2^* and Y_3^* in Model 1 should be considered as measuring two dimensions and the Y_3^* should not be removed or combined with Y_2^*.

From the above results, we re-examined the meaning of ”N/A” responses from the MCA scale. For the faculty sample, even from the ”N/A” response patterns (Table 3.1), we noticed that the percentages of ”N/A” responses were essentially larger. Considering that faculty participants were rating themselves, the response patterns might indicate that participants were certain if an item was applicable to their own conditions. We also found consistent results in the two IRTree model, where there were small, but non-ignorable negative correlations between the nodes Y_1^* and Y_2^*.

The negative correlations indicated that faculty who tended to choose the ”N/A” option were more likely to rate their mentoring skill levels higher; furthermore, we could interpret this phenomenon from a different perspective – when faculty participants were certain that the measured skills were applicable to them, they rated high on those applicable measured skills. ”N/A” responses could somewhat indicate the level of certainty or confidence; when the confidence level was low on an item, individuals might choose ”N/A” over a low rating. With this in mind, even the AIC and BIC indicated that the unidimensional IRT model was better than the IRTree models, we should be cautious about model selection, since the assumption of treating ”N/A” responses as missing data might not hold.

The results from the two IRTree models using the student sample yielded two different suggestions for handling ”N/A” responses. In Model 1, $\sigma_{Y_1^* Y_2^*}$ was close to 0, so we concluded that the nodes Y_1^* and Y_2^* were independent and the ”N/A” responses were random to the future response process. In this case, we could treat the ”N/A” responses as missing at random. In Model 2, we noticed that $\sigma_{Y_1^* Y_2^*}$ was negative and non-ignorable. This indicated that students who tended to choose ”N/A” responses were more likely to rate higher on their faculty mentors’ skill levels.

When student participants rated their mentors, although rating others, instead of self-rating encountered less of the ”confidence” issue, they might face a similar issue. Students rated the mentoring skills that they received, which might not reflect the actual applicable skills to their mentors. Students might only rate partial skills that they experienced. If their mentors were not skilled in some aspect and as a result, the mentors never expressed those mentoring skills, students could only claim that they did not have the related experience and declared that the measured skill aspects were not applicable. In an extreme case scenario, if students only experienced the mentoring aspects that their mentors were extremely skilled, reflecting on the MCA survey response scoring, students probably rated higher than their mentors’ self-rating, due to a lack of information in the mentoring aspects they did not experience. This hypothesis could also explain the observed ceiling effect we mentioned in Chapter 2. Similar to the previous conclusion, we should be cautious about model selection based on AIC and BIC, because the ”N/A” response effect could not be fully eliminated.

CHAPTER 4

Evaluating the Impact of the BUILD Scholar Program on First Year College Students' Intent to Pursue Science-related Research Careers

4.1 Introduction

Given the high demand for talent in STEM careers, especially in Medical Science and Biomedical Engineering, the Department of Education has urged all young people to "be prepared to think deeply and to think well so that they have the chance to become the innovators, educators, researchers, and leaders who can solve the most pressing challenges" (2015). To encourage undergraduates to enter the biomedical research field, the National Institute of Health (NIH) has provided opportunities for underrepresented minority (URM) students to participate in research training.

Studies showed that undergraduate diversity training programs, such as the NIH-funded Bridges to the Baccalaureate (B2B) program (n.d.), the Research Training Initiative for Student Enhancement (RISE) program (n.d.) and the Maximizing Access to Research Careers (MARC) award (n.d.), could increase the likelihood of URM students expressing intentions and then following through on those intentions to pursue careers in science-related research. (Schultz et al., 2011; MacLachlan, 2012). In 2013, the NIH started a new set of initiatives using a transformative approach to "supplant less-effective practices and methods to have a broad and sustained impact on the diversity of the NIH-funded biomedical research workforce" (Funding Opportunity Announcement, 2013a, 2013b, 2013c). The overall goal is to "[promote] diversity in the NIH-funded biomedical, behavioral, clinical, and social sciences (collectively termed 'biomedical') research workforce" (Funding Opportunity Announcement, 2013b). The funding provides constant support for individuals from diverse backgrounds who are underrepresented in biomedical research. Participants can receive training and mentorship throughout their undergraduate and graduate education and into their early career. With this support, the program could contribute to diversifying the candidate pool in biomedical research at different educational stages. This funding opportunity resulted in the activation of the Diversity Program Consortium (DPC), a collaborative program with NIH, consisting of the BUilding Infrastructure Leading to Diversity (BUILD) Initiative, the National Research Mentoring Network (NRMN) Initiative, and the Coordination and Evaluation Center (CEC).

4.1.1 BUILD Scholar Program

The BUILD initiative is a set of NIH-funded experimental training programs that aims to attract undergraduate students from diverse and historically excluded backgrounds into biomedical research fields and through innovative methods, prepare them for academic success and career readiness (McCreath et al., 2017). The BUILD initiative is designed to explore the most effective ways to engage students from underrepresented backgrounds in biomedical research, helping them progress on the path to becoming future contributors to the biomedical research fields. The BUILD primary sites each set up their own selection process and structure and facilitate their BUILD programs according to the characteristics of their specific student population. In most of the sites, BUILD programs have unique names associated with the BUILD interventions at their sites. Each BUILD site submitted their individual proposals and received funding to implement their own, local and specific interventions. The 10 successful BUILD sites received their funding through a request for proposals from the NIH in 2014. Beginning in the fall of 2014, the first wave of funding included more than $500 million allocated to these 10 sites, a national mentorship network, and a coordination and evaluation center that collectively formed the DPC.

The BUILD scholar program is one of the most intensive BUILD programs, and participants receive support, varied in both nature and scope, related to research training, academic support, financial aid and scholarships, professional development, extensive advising and mentoring, and so forth. Four out of the 10 BUILD scholar programs accept newly admitted first-year students. In order to protect identities, the four primary BUILD sites were referred to as Sites A, B, C and D. Table 1.1 (adapted from Davidson et al., 2017, p. 166) summarized the basic institutional information of the BUILD primary sites prior to the start of the BUILD program.

Table 4.1: BUILD Scholar Financial Support

	Financial Support*			**Program Duration**
	Tuition and Fees	**Stipend**	**Other Funding**	
Site A	15 units/semester	$1114/month	Publication costs; Travel awards	Renewable annually
Site B	Fully covered, up to 30 credits/yr	Not specified	Program award up to $5000/yr; Travel awards	Renewable annually, up to 4 years
Site C	Tuition support	Monthly stipend	Research awards; Travel awards	Structured 2 years
Site D	Up to 30 credits/yr	$1114/month	Travel awards	Renewable annually, up to 3 years
NIH TL4**	60% of tuition, up to $16000/yr	$1114/month ($13368/yr)	-	-

*The amount of financial support was reported or estimated based on the aids amount in FY 2020

**The listed TL4 financial support amount was for the Freshman/Sophomore career level in FY 2020

The BUILD scholars at these four sites receive no less than the NIH-defined TL4 (NIH: NOT-OD-20-070, 2020) financial support (Table 4.1), and engage in research enrichment and professional development activities (Table 4.2). Table 4.1 presents the site-specific financial support for first-year BUILD participants in Fiscal Year (FY) 2020 and the possible program duration. We collected the information through the BUILD program websites and the NIH announcements. Table 4.1 indicates that the BUILD scholar program financially supports its participants to a large extent, since it covers the majority of their tuition and fees and provides monthly stipends as well as other program-related awards. Prior to FY 2020, qualified Freshmen/Sophomores and Juniors/Seniors received stipends on two different levels. Freshmen/Sophomores received about three quarters of the amount that the upperclassmen were awarded. All BUILD scholars joining the program on and after the FY 2020 received the same level of stipend, which was similar to the previous level for Juniors/Seniors.

Table 4.2: BUILD Scholar Research Enrichment and Professional Development Activities

Activities	**Site A**	**Site B**	**Site C**	**Site D**
Program entry points	Fr/Sph/Jr/Sr	Fr/Transfer (Sph/Jr/Sr)	Fr/Jr (transfer)	Fr/Sph/Jr
Summer bridge	O	C	C	O
Learning community	C	C	C	-
Enrollment in novel curriculum	C	C	C	C
Research training (mentored)	C	C	C	C
Undergraduate research experience	C	C	O	C
Conferences (local/national)	O	C	C	C
Career advancement & development	C	C	C	C
Other funding support	O	O	O	O

Abbreviations: C = Compulsory; O = Optional; - = Not mentioned

Table 4.2 reported program entry points and the main program activities offered to the BUILD scholars at each site. These data were collected from programs' internal reports and participation data. Table 4.2 showed that apart from learning community activity which was not offered by Site D, other activities were offered either as compulsory or optional at all four sites. Activities such as novel curriculum, mentored research training, as well as career advancement and development were compulsory at all four sites. Table 4.2 reflected the fact that although programs were designed and facilitated by each site, the main activities were similar across sites.

To participate in the BUILD scholar program as a first-year college student, candidates need to apply to the program during the spring semester of their last year of high school. The BUILD scholar program staff review applications and select suitable candidates based on site-specific criteria (Table 4.3). Table 4.3 showed that almost all sites required candidates to be full-time students, in biomedical related majors, and holding relatively high GPA. They also preferred their BUILD scholar candidates had interests in science and research, planned to obtain graduate degrees, and intended to pursue biomedical related careers.

Table 4.3: BUILD Scholar Program Selection Criteria

Selection Criteria	**Site A**	**Site B**	**Site C**	**Site D**
Status	Full-time	Full-time, first-year	Full-time, first-year	Full-time, first time college student
Major	Biomedical	Biomedical, psychological, behavioral or social sciences	STEM related to biomedical science	Approved BUILD majors
Academic	2.75 GPA	3.0 GPA	3.0 GPA	High school GPA, SAT/ACT
Intended Career	Careers in healthcare, biomedicine; research in life, social sciences		Careers in biomedical or behavioral research fields	Research scientists in biomedical fields
Expected Degree		Grad-level, especially doctoral studies	STEM related graduate school education	Graduating with a Ph.D.
Science & Research Interests	Interests in biomedical or health research	Desire to learn about conducting formal research and engaging in real-world projects	Documented interest in research in the biomedical or behavioral sciences	

Although the BUILD scholar program, like many other NIH-funded diversity training programs, does not have standardized interventions and selection criteria across sites, the four sites that allow first-year student participation share similar program activities (Table 4.2) and selection criteria (Table 4.3). In addition, the programs at BUILD sites share common goals, such as stimulating students' interests in science and increasing their intent to pursue science-related career paths.

4.1.2 Purpose of the Study

Research has shown that college students' intentions to pursue a biomedical science research career are often significant and strong predictors of subsequent pursuit of such a career (Young, Fraser, & Woolnough, 1997; Pascarella & Staver, 1985; Dibenedetto, Easterly, & Myers, 2015; Sahin, Ekmekci, & Waxman, 2017); therefore, it is reasonable to expect that college students' declared career intentions could serve as proxies or early indicators for future career choices (Ajzen, 2011; Deci & Ryan, 1985). Given that one of the core goals of the BUILD initiative is to diversify the biomedical scientific workforce, an early indicator of the program's success can be operationalized through students' expressed intentions to pursue science-related research careers. Examining early career intentions among first-year college students as an evaluation of the BUILD initiative's initial efficacy also provides formative feedback to funders and program directors; this early feedback is especially critical as an evaluation of whether the program actually increases individuals' likelihood of pursuing such careers could span well over a decade – from program participants' matriculation and completion of graduate degrees to subsequent entry into the scientific workforce.

The purpose of this study is to examine the effectiveness of an undergraduate diversity training program, the BUILD scholar program, on students' intent to pursue science-related research careers during their initial stage in college, and the corresponding research question is: Does participation in the BUILD scholar program during the first year of college impact students' intent to pursue science-related research careers? This study relies primarily on longitudinal survey data and programmatic administrative records to examine the effectiveness of the BUILD intervention on increasing the likelihood of participants' intentions of pursuing biomedical research careers. When combining these two data sources, we potentially construct a quasi-experimental design. Methodologically, this study aims to demonstrate an approach of using program data and self-reported data from surveys to explore potential causal relations, and then applying multi-stage matching and sensitivity analysis to prevent the confounding with the program effects and to examine the plausibility of threats to internal validity.

4.1.3 Significance of the Study

As a part of the BUILD program evaluation, this study provided credible evidence of the effectiveness of the scholar program on increasing first-year college students' intent to pursue science-related careers. The findings can be generalized to first year college participants in other federally funded initiatives designed to contribute to the diversity of undergraduate and graduate science education as well as the scientific workforce. Much of the research in this area has focused on the experiences of juniors and seniors who participated in mentored research experiences; therefore, this study's focus on first-year students expands our understanding of the effectiveness of similar initiatives targeting first-year college students.

Using a quasi-experimental design, this study aims to examine the causal influence of a federally funded intervention program on changes in first-year college students' likelihood of intending to pursue biomedical science research careers. Many studies that proved the positive influence of the diversity training programs for enhancing undergraduate students' intent to pursue science-related research careers were based on anecdotal or correlational evidence. We attempt to go beyond the correlational relations. Our study analyzes the BUILD scholar program's effects on an intended program outcome variable, using analytical approaches that provide a reference for examining program effectiveness on other program-intended outcome variables. Furthermore, we hope that this study presents a useful example for researchers who are interested in exploring causal relations using data with similar components.

4.2 Review of Intent to Pursue Science-related Research Careers

"Roads diverged in a wood, and I —
I took the one less traveled by,
And that has made all the difference."

— Robert Frost

An important goal of higher education is helping students achieve their desired outcomes, which includes helping students find and pursue their intended career paths (Peterson, 1993; Brown, Glastetter-Fender, & Shelton, 2000; Gianakos, 1999). For college students, the decision to pursue a particular life-long career is one of the most important and most difficult self-selections (Freedman, 1999; Peterson, 1993; Willis & Rosen, 1979). Research shows that college students' career choices can be approximated by their declared career intent (Ajzen, 2011; Deci & Ryan, 1985). Students who have ambitions for working in science-related research fields while in college are typically more likely to find themselves in science-related research careers compared to their college peers who do not share these goals (Young et al., 1997; Pascarella & Staver, 1985; Dibenedetto et al., 2015; Sahin et al., 2017). Factors that influence students' intent to pursue science-related careers include demographic variables, pre-college academic performance and experience in STEM education, self-efficacy in science and research, and college experience (Crisp, Nora, & Taggart, 2009; MacLachlan, 2012; Sweeney & Villarejo, 2013; Pascarella & Staver, 1985; Young et al., 1997; Bottia, Stearns, Mickelson, Moller, & Parker, 2015; Wang, 2013; Sahin et al., 2017). Particularly, undergraduate diversity training programs have proved to be effective for increasing students' intent to pursue science-related careers, especially for the URMs; however, there is a lack of research quantifying program effects on first-year college students (MacLachlan, 2012; Hurtado et al., 2008).

4.2.1 College Students' Career Development: From Intent to Actions

Though not specifically studying college students' career development, many career development theorists have described how young adults seek career interests. For example, Super's developmental self-concept theory (1953) identified by age groups five distinct career developmental stages. Most college students belong to the age group of 15-24, which Super's theory identifies as the "Exploration" stage. The key words and phrases of this stage include "trying things out, crystallizing, specifying, and implementing career choice" (Super, 1990). In Super's theory, the career development process is driven by one's self-concept (1990). Holland's career typology theory of vocational behavior (1959) states that an individual career choice is determined by the personality type they most resemble. In Holland's theory, personality types are characterized by interests, preferred activities, beliefs, abilities, values, and characteristics. Social cognitive career theory (Lent, Brown, & Hackett, 1994; Lent, 2005), which is an extension of social cognitive theory into career development theory (Lent & Brown, 2006), describes that individuals' beliefs largely influence their career choices. These theories, as well as many other career development theories, are similar in two aspects. One, they acknowledge the influence of external or environmental factors on the internal factors, and in this way career development seems to be a constructive process. Two, all these theories emphasize that internal factors, (i.e., self-concept, self-efficacy, beliefs, motivations, or interests), are the major forces driving career choices. As the saying goes, "where there's a will, there's a way."

In science education literature, other theories, such as the theory of planned behavior (Ajzen, 2011), self-determination theory (Deci & Ryan, 1985), and flow theory (Csikszentmihalyi, 1997) have been used as frameworks to explore students' persistence in science education and the pursuit of science-related careers. As a result, inner motivations and early aspirations serve as the primary factors influencing the career choices of college students (Ajzen, 2011; Deci & Ryan, 1985; Csikszentmihalyi, 1997). For example, Mishkin, Wangrowicz, Dori, and Dori (2016) studied the career choice of undergraduate engineering students under the theory of planned behavior, and they found that students who expressed positive attitudes towards engineering education were less likely to be influenced by external factors such as subjective norms. Similarly, Lavigne and Vallerand (2010) developed the hierarchical model of intrinsic and extrinsic motivation (HMIEM) in science education based on the self-determination theory, and identified "a conscious internalization of personal valuable beliefs" as the highest level of self-determination (p. 2344). Ellwood and Abrams (2018) concluded that experiences of flow could sustain students' motivation in science education and elevate achievement outcomes. In their study, students' motivation and achievement outcomes were inseparable.

Empirical studies suggested that students' intent to stay on the science track was one of the most influential factors determining whether they remained in the science related fields (Sweeney & Villarejo, 2013; Bottia et al., 2015; Wang, 2013). Many more studies directly used students' declared intent or desire as the representation or approximation of the likelihood that students would pursue science-related paths (Young et al., 1997; Pascarella & Staver, 1985; Dibenedetto et al., 2015; Sahin et al., 2017). Theoretical and empirical evidence suggested the potential for using students' intent to pursue science-related research careers as an approximation of their future actions in pursuit of science-related research careers. Furthermore, efforts to increase students' intent to pursue science-related research careers could help students get onto science-related paths (Sahin et al., 2017).

4.2.2 Influential Factors of College Students' Science Career Choices

In career development theories, influential factors of career exploration and career tendency development can be conceptualized by a set of internal motivations and ambitions and a set of external factors related to school context, parental influences, and social networks (Duffy & Sedlacek, 2007). Internal factors include self-efficacy (Betz, Klein, & Taylor, 1996; Gianakos, 1999; Peterson, 1993; Willis & Rosen, 1979), psychological development (Harren, 1979), outcome expectations (Lent, Brown, & Hackett, 2000), developmental trajectory (Super, 1980), personal environment fit (Holland, 1997), parental influence (Fisher & Padmawidjaja, 1999; Hartung, Lewis, May, & Niles, 2002), and social cognition (Lent et al., 2000). External societal factors (or rather societal factors) include social stratification (Anctil, Hutchison, & Smith, 2013), subjective norms (Ajzen, 2011), social demand (Willis & Rosen, 1979), and economic climate (Stone, Van Horn, & Zukin, 2012). Educational studies have usually focused more on internal factors as well as external factors associated with educational equity and inclusion, such as social stratification and subjective norms. In science education literature, factors that influence college students' persistence on science paths and their choice of science-related careers can be summarized into four categories: demographic variables, pre-college academic performance and experience in STEM education, self-efficacy in science and research, and college experience. Factors under these categories somewhat reflect those abstract factors in career development theories.

4.2.2.1 Demographic Variables

In educational studies, researchers often report students' demographic information or personal background variables such as race/ethnicity (Sweeney & Villarejo, 2013; Crisp et al., 2009; Herrera & Hurtado, 2011) (MacLachlan, 2012), gender identity (Mishkin et al., 2016; Pascarella & Staver, 1985; Sweeney & Villarejo, 2013; Riegle-Crumb & Morton, 2017; Robnett, 2013; Amelink & Creamer, 2010; Sahin et al., 2017), social economic status (Crisp et al., 2009; MacLachlan, 2012) and parents' education (Sweeney & Villarejo, 2013; Pascarella & Staver, 1985), in educational studies. Many researchers use these demographic variables to categorize the underrepresented minorities (URMs) in higher education, and to explore differences in educational outcomes due to the underrepresentation. Their findings are similar – URMs in STEM education (e.g., students of color, female students, students from low-income families and first-generation college students) persist at lower rates in the STEM fields than their well-represented peers (Crisp et al., 2009; Mishkin et al., 2016; Pascarella & Staver, 1985).

The NIH's identification of the URMs in biomedical research aligns with the previously mentioned literature. Maccalla et al. (2020) reviewed science education literature and NIH guidelines and summarized the identification of URMs through item responses of race/ethnicity, gender identity, social economic status, parents' education, and many other variables. They noted that URM identification could be limited to the availability of information (e.g., appearance of items in the survey), and should be evaluated according to research purposes and actual situations.

4.2.2.2 Pre-college Academic Performance and Experience in STEM Education

Many studies proved that pre-college academic performance and experience in STEM education were usually good predictors of college students' intent to pursue science-related careers (Pascarella & Staver, 1985; Young et al., 1997; Bottia et al., 2015; Wang, 2013; Crisp et al., 2009; Sahin et al., 2017). These variables can be pre-college academic aptitude (Pascarella & Staver, 1985; Crisp et al., 2009), high school STEM learning experience (Bottia et al., 2015), high school math achievement and math self-efficacy (Wang, 2013; Sahin et al., 2017), high school percentile (Crisp et al., 2009), etc.

4.2.2.3 Self-efficacy in Science and Research

As mentioned in the career development theories, self-efficacy is an important internal driver of career choice. Self-efficacy describes individuals' confidence about their competence in a domain field (Bandura, 1991). Pajares and Schunk (2001) suggested that "individuals tend to engage in tasks about which they feel competent and confident." Students' self-efficacy influences their decision making; they pursue what they believe they can succeed at and then increase their efforts. The domain-specific self-efficacy in pursuing science-related research careers are described as science identity and researcher self-efficacy.

Science identity. Science identity indicates "the extent to which students conceive of themselves as scientists" (CIRP Constructs, n.d.). Estrada, Woodcock, Hernandez, and Schultz (2011) observed that underrepresented students' sense of self-efficacy was significant measure of their ability to integrate into an academic social system in science, and perhaps more importantly, their personal identification as a scientist had a stronger relation to their persistence in science. Carlone and Johnson (2007) found that by increasing students' tendencies to feel, think, behave, and be recognized by meaningful others (e.g., faculty role models) as a "science person," URMs had a much greater chance of believing in their ability to succeed in science.

Researcher self-efficacy. Researcher self-efficacy, research self-efficacy, or scientific research self-efficacy describes "students' sense of confidence to engage with the scientific method" (CIRP Constructs, n.d.). Adedokun, Bessenbacher, Parker, Kirkham, and Burgess (2013) grouped students' intention to pursue careers in STEM under the construct of aspirations for research careers, exploring the relations among research skills, research self-efficacy, and student aspirations for research careers. The results suggested that researcher self-efficacy is a predictor of student aspirations for research careers.

4.2.2.4 College Experience

Sweeney and Villarejo (2013) believed that external factors associated with students' college experience, such as research experience, coursework, peers, and mentors, influenced college students' science-related career choices. Schultz et al. (2011) specified that undergraduate minority training programs, the presence of a scientific mentor, and research experience moderated the decline of students' intent to pursue a research career across their college years. Pascarella and Staver (1985) found the positive influence of on-campus work in science on science career choices. Crisp et al. (2009) reported that initial college experiences in STEM education, such as the first-semester academic performance and introductory science course-taking sequences, were associated with the likelihood of earning a STEM degree. Wang (2013) found similar results that students' initial post-secondary experiences were related to choosing a STEM track. Particularly, STEM major intent, academic interaction, receiving financial aid, and expecting to earn a science graduate degree were positively associated with entering into STEM fields.

4.2.3 Enhancing URMs' Science Career Intent Through Diversity Programs

For URM students, the persistence of the intention to pursue a career in science is especially critical. Since the 1970s, despite a convergence in intentions among URM and white students majoring in a STEM-related field when initially enrolling in college, disparate completion rates between URM students and their well-represented peers have persisted (Rask, 2010). An URM student can face substantial difficulties when attempting to complete their STEM degree. Gibbs Jr and Griffin (2013) noticed that in biomedical sciences, there were significant amounts of variance in the career choices between the two populations at the undergraduate level, and even at the graduate level, non-URMs saw greater freedom to pursue their interests. The extra challenges URMs face while pursuing a scientific research career call for support and research.

As mentioned in the previous section, college experiences, such as research experience, financial aid and academic interaction often positively influence students' intent to pursue science-related careers. These factors are often included in an undergraduate diversity training program. Researchers found that diversity training programs, such as Bridges to the Baccalaureate (B2B), Research Initiative for Scientific Enhancement (RISE) and Maximizing Access to Research Careers (MARC), as well as other science training activities, helped URMs enter science-research tracks and increased their intent to pursue science-related careers (MacLachlan, 2012; Schultz et al., 2011; Crisp et al., 2009; Pascarella & Staver, 1985; Sweeney & Villarejo, 2013; Dibenedetto et al., 2015). Even under the condition that students' interests in science tended to decline across their college years (Hurtado et al., 2008), students supported by diversity programs had a higher probability of keeping their intent to pursue science-related research careers (Schultz et al., 2011).

Although considerable research has shown the positive influence of undergraduate diversity training programs for enhancing URM's intent to pursue science-related research careers, most of the conclusions were drawn from anecdotal or correlational evidence. There was a lack of literature that clearly identified causal effects of a diversity training program on students' persistence in science careers.

Additionally, as MacLachlan (2012) mentioned, the participants in most undergraduate diversity training programs were already likely to be high achievers in science research due to the program design or selection criteria. They might have completed introductory gatekeeper science courses, maintained a high college GPA, and been exposed to certain research training before joining the programs. As a result, first-year students are rarely involved in such programs (Hurtado et al., 2008). This implies that many research findings in this topic drew from programs that only admitted upperclassmen in college. Hurtado et al. (2008) recognized the importance of "early efforts to provide structured opportunities for students" and called for future research exploring diversity program effects on first-year students.

4.3 Theoretical Framework: Evaluating BUILD Program Effects

The standards of experiments are unlikely to be achieved by design in educational studies because it is often unethical to randomly assign students to a treatment or control group. Quasi-experiments can be good alternatives, for they share similar features with the experiments, apart from the random assignment. Shadish, Cook, and Campbell (2002) commented on the primacy of control by design in quasi-experimental research and suggested that "the usual alternative to design controls are statistical controls that attempt to remove confounds from effect estimates using statistical adjustment after the study is done" (p. 105). This indicates that even without an experimental design, certain levels of design controls combined with proper post-hoc statistical controls could lead to a reasonable causal claim.

We developed the theoretical framework of evaluating BUILD program effects based on this idea. In this section, we will go over the theoretical background of causal relations, review previous attempts to examine causal relations in science-related interventions and students' science-career intent, and using the scholar program as an example, introduce feasible approaches for evaluating BUILD program effects. We avoided using complex mathematical expressions in this section in the hope that this paper could help a broad range of researchers interested in examining causal program effects to construct their analyses. We did not intend to give a holistic review of the Rubin causal model; instead, we provided a showcase of evaluating educational program effects.

4.3.1 Causal Effects

Our discussion is tailored to educational programs. Following Rubin (1974), we assume that 1) we are able to know when and who received the intervention (or did not receive it), and 2) the membership of being in the treated group and that of being in the control group are exclusive of each other. Under the random assignment of intervention (or selection independence), the causal effect is defined as the difference between the average post-intervention measure of the treated group outcome and that of the control group outcome. We also assume that the stable unit treatment value assumption (SUTVA, see definition on Imbens and Rubin, 2015, p. 10) holds and there is no spillover.

A causal relationship should follow at least three requirements: 1) the cause precedes its effects, 2) the cause covariates with its effects, and 3) alternative explanations are implausible (Shadish et al., 2002). For educational programs with manipulated interventions, temporal relations and correlations among variables are easy to identify. In this case, examining alternative explanations is the key to determining causal relations. We use D, Y, and X to represent treatment assignment and outcome variable and covariate, respectively. Under random assignment, X and D are independent, and the treatment effect is the difference in the observed outcomes (Figure 4.1a). When X is associated with D, meaning that random assignment is not feasible, X is covariate with D as well as Y (Figure 4.1b). In this situation, we usually describe X as a confounder, which can be observed or latent. Confounders are those potential alternative explanations that we need to carefully examine. After we address the confounder issues, we can estimate the causal effects from the differences in the observed outcomes. The treatment effects are identifiable when the conditional ignorability holds, which means among units with the same X, treatment D is as good as randomly assigned. In this process, commonly used methods include matching, weighting, regression and sensitivity analysis for unobserved confounders.

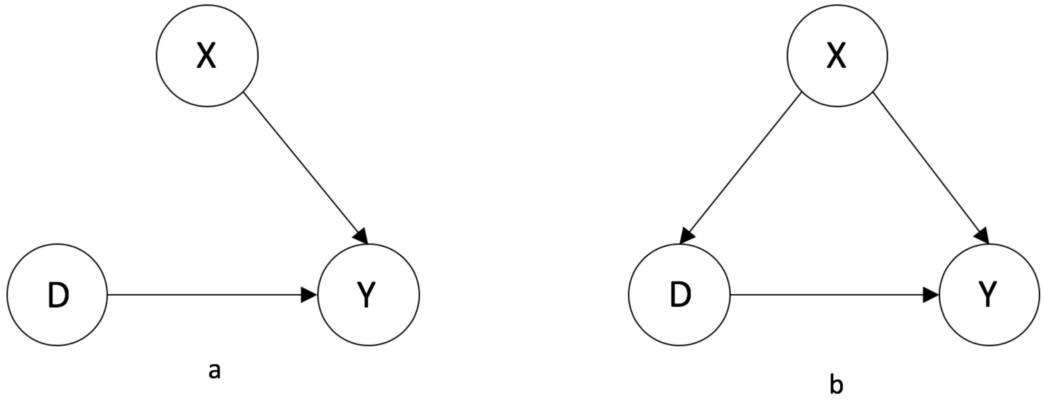


Figure 4.1: Simple Illustration of Confounders

4.3.2 Previous Attempts

Previous studies related to program effects and science career intent rarely attempted to report causal relations. In several existing studies, we see the efforts of using statistical tools to extract possible causal effects. For example, Pascarella and Staver (1985) explored the causal effect of on-campus work in science on college students' science career choices, using the input-environment-output (IEO) framework (Astin, 1970a). In their last model, they used partial correlation to control for the input bias. However, Pascarella and Staver (1985) ignored the temporal relations among input and environmental variables, nor did they sufficiently classify variables into inputs or environment due to their lax research focuses. Though their analysis was somewhat limited by the statistical approaches proposed by Astin (1970a, 1970b), their study was still undeniably valuable. Pascarella and Staver (1985) inspired us to consider the possibility of using the IEO framework to explore the effects of a well-defined environmental variable.

In a longitudinal study to explore the program effectiveness of the RISE program, Schultz et al. (2011) put great effort into constructing causal relations from observational data. In this study, the participants were not randomly assigned into the treatment or control groups. Researchers implemented a common strategy, propensity score matching (PSM), to address the covariance imbalance between the treated group and control group at the baseline. The logic of their analytical design was rigorous; however, their results were somewhat problematic for two reasons. Firstly, they only reported the balance of the propensity scores after matching and did not report the balance of the covariance used for PSM. This left uncertain whether after matching, the treatment and control group achieved similar covariance distributions. Judging from the large difference of the treatment indicator estimates between their model 4 and model 5 (Schultz et al., 2011, p. 104, Table 3), the matching or intervention indicator (β_{30}) and the propensity score (β_{01}) seemed to be associated, which indicated that their after-matching balance was questionable. Secondly, researchers matched cases across sites in a multi-site study without any adjustment. Nevertheless, Schultz et al. (2011) demonstrated the use of matching to achieve conditional ignorability.

Our goal was to identify an approach that could help us maximize within group matches. To account for the selection differences at local sites, Rickles and Seltzer (2014) proposed a two-stage propensity matching strategy (2SM). They firstly identified matched cases at local level based on propensity scores, and for those that failed to find a close match at local site, they found matched cases across sites using Mahalanobis distance. After matching, they adjusted the outcome values for those matched cases from non-local sites. The general goal was to ensure the similarity of the treated and control groups in a multi-site study. The downside of matching was that it could jeopardize sufficient overlapping of the covariates and common support, because not for any value of X, the unit could have received treatment or control. Another strategy was moving forward directly to the regression analysis with covariates in the regression model, and checking backwards on the impacts of covariates to estimate if the treatment was effective under the influence of the covariates. This approach could be realized by the sensitivity analysis (Cinelli & Hazlett, 2020), which not only helped with assessing the impacts of observed confounders, but provided information regarding whether unobserved confounders could influence the results as well.

4.3.3 Evaluating BUILD Program Effects

Inspired by Pascarella and Staver (1985), we designed our research framework based on the IEO framework (Astin, 1970a, 1970b; Astin & Antonio, 2012). Although viewed as a mediation model in many applications, the IEO framework (Figure 1 in Astin, 1970a, p. 225) is more similar to causal inference under selection on observables, if considering the environmental variable to be the treatment and input variables to be observed confounders (Figure 4.2). As a matter of fact, in his original IEO papers, Astin (1970a, 1970b) mentioned confounding issues several times, although his attention focused on multiple relations among the three components. Astin and Antonio (2012) stated that the focus of the IEO framework was on the environmental effects on educational outcomes, and they described the environmental variables as treatments, interventions, programs, etc. From this perspective, the IEO model could be considered an applied version of a causal inference framework in educational studies, with the assumption that inputs naturally confound with environment and output. Astin, possibly influenced by his training in psychological and educational measurement, was a pioneer of using a neat diagram (Figure 1, 1970a, p. 225) to illustrate causal relations and confounding issues in educational studies. The IEO illustration (1970a) encapsulated the details in the path analysis (Wright, 1921) – the latter was considered as the "direct ancestor" (p. 11) of statistical methods for analyzing causal effects (Pearl & Mackenzie, 2018) – and created a general framework for analyzing educational causal relations from observational data. Astin's IEO framework (1970a, 1970b) was a little ahead of its time, several years before the maturity of the statistical framework proposed by Rubin (1974) and Cochran and Rubin (1973). Interestingly, although both frameworks were more or less influenced by D. T. Campbell and his colleagues' earlier work in quasi-experimental evaluations (e.g., Campbell & Stanley, 1963; Campbell & Erlebacher, 1970), they were not interacted much during the past half century. Combined with the statistical techniques of causal inference analysis, the IEO framework could go beyond correlations and bounds to become a more powerful tool for analyzing causal inference in educational studies.

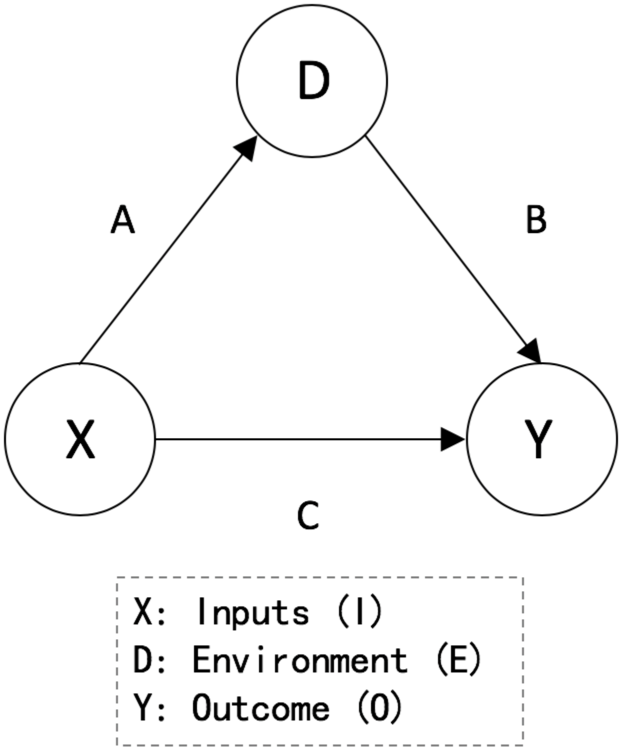


Figure 4.2: Evaluating Causal Effects in Higher Education

The general framework we proposed for evaluating BUILD program effects (Figure 4.3) was similar to the illustration of the IEO framework. Our focus was similar to the IEO framework in that we wanted to know the effectiveness of an educational intervention on a specific outcome after controlling for demographic and pre-college inputs that influenced the outcome as well as selection into either the treatment or control conditions of the intervention. This encourages researchers to examine the mitigating effects of intervention efforts and policy implementation of key demographic characteristics to provide greater insight into the effectiveness of these strategies to achieve more equitable outcomes in education.

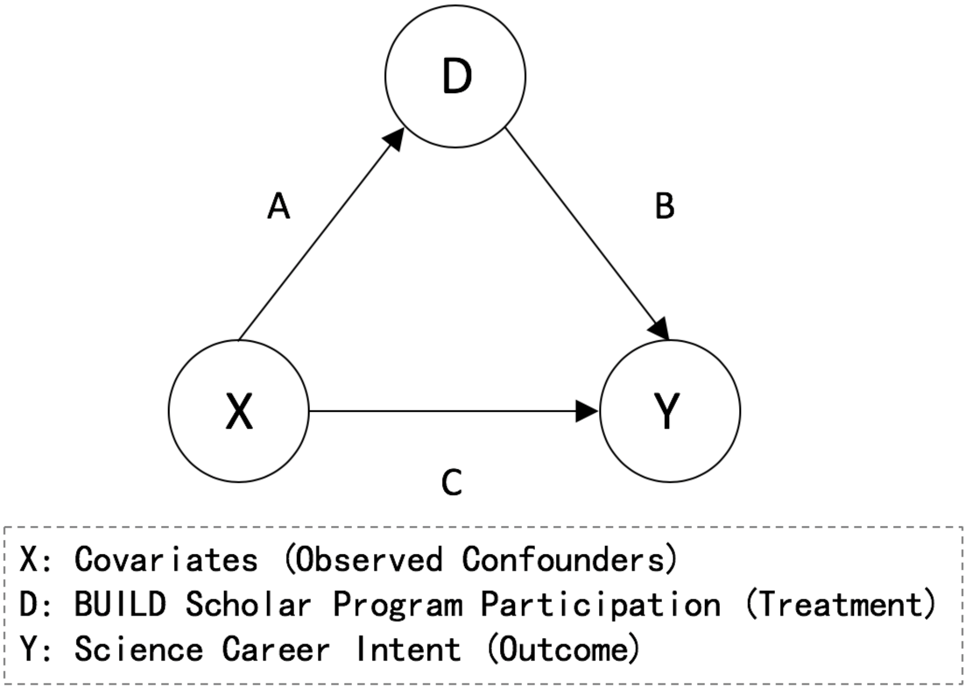

D
A
B
X
Y
C
X: Covariates (Observed Confounders)
D: BUILD Scholar Program Participation (Treatment)
Y: Science Career Intent (Outcome)

Figure 4.3: Framework of Evaluating BUILD Program Effects

After considering the study variables and learning from the previous attempts, we identified participation in the BUILD scholar program as the intervention (D) or cause and the students' post-test responses of intent to pursue science-related careers as the specific outcome (Y). According to the program description, there were, by design, selection criteria in each site (see Table 4.3), which indicated the existence of site-level and/or cohort-level selection bias (observed confounders). We have also learned from previous literature that students' intent to pursue science-related careers can be influenced by demographic variables, pre-college academic performance and experience in STEM education, and self-efficacy in science and research. These variables, however, might also be associated with the likelihood of being selected into the BUILD scholar program. This information helps us identify potential observed confounders, which were included among the confounders represented by X in the model depicted in Figure 4.3.

With a designed environmental variable (i.e., the BUILD scholar program), our framework and potential analytical approaches were straightforward. The study had a primary focus on the magnitude of the effects on path B (in Figure 4.3) and the plausibility of this quantity. We either found a way to "disconnect" path A before conducting analysis for estimating the magnitude of effects of path B, or we employed a certain approach to estimate the impact of path A on changing the magnitude of effects on path B. In addition, as much as we tried to include all possible covariates, we were not sure if any unobserved confounder could threaten the plausibility of our analysis. Accordingly, we took two approaches to estimate potential program effects and introduced the suitable conditions in the next section. In the first approach, we applied a multi-stage matching strategy similar to the 2SM approach (Rickles & Seltzer, 2014) to ensure the largest extent of local similarity. After examining the covariance balance, we used regression to estimate program effects. In the second approach, we skipped the matching and applied the regression analysis on the whole dataset. After getting the regression results, we performed sensitivity analysis to examine the plausibility of the program effects and the influence of potential confounders.

4.4 Methodology

In this study, we used BUILD program participation data and longitudinal survey data to examine the effectiveness of the BUILD scholar program on students' intent to pursue science-related research careers during their initial stage in college. To address the research question, “Does participation in the BUILD scholar program during freshman year impact students’ intent to pursue science-related research careers?” we identified 4 primary BUILD sites that provided the scholar program interventions for their first year students, and based on the program characteristics, we constructed potential causal relations. Our goal is to realistically assess the BUILD scholar program effects and we hope that our demonstration of analytical approaches can be an example for future research. In this section, we will introduce the data sources of the study, the sample, the process for variable construction, as well as the analytical approaches and their conditions and limitations.

4.4.1 Data Sources and Sample

The BUILD sites reported program participation data through an internal tracker, which became the data source of BUILD program participation. Through the program participation data, we identified who participated in the BUILD scholar program and when they joined and left the program. In this study, we exclusively selected students who joined the BUILD scholar program during their first (usually fall) semester in college and who were still enrolled in the program in the following (usually spring) semester. We linked the participation records to longitudinal surveys to construct datasets for analysis.

Data used in this study were majorly collected through the Higher Education Research Institute’s (HERI) Cooperative Institutional Research Program (CIRP) Freshman Survey (TFS) and the NIH Diversity Program Consortium Student Annual Follow-Up Survey (SAFS). The TFS, originally developed by HERI in the 60s and revised based on historical trends in the national sample data and issues currently facing higher education institutions, was administered to incoming first-year students in the participating institutions before they started their classes in the first semester, and was open for incoming first-year students to respond usually from late March till early October (CIRP Freshmen

Survey, 2021). Generally, the TFS was distributed to the total sample of incoming first-year students, and for the studied BUILD sites, the response rates approximately ranged from 20% to 50% across sites and cohorts since the beginning of the BUILD program. The response rates for freshman year BUILD scholar program participants were much higher, usually over 90%.

As a follow up to the HREI Freshman Survey, the SAFS was administered to continuing undergraduate students in the BUILD sites, and was open to respond from mid-spring till early summer. We used the HERI CIRP student survey items (CIRP Constructs, n.d.) as the initial item pool from which to draw items to include on the SAFS. We revised items and tailored the content to fit the purposes of BUILD program evaluation. To reduce the administrative burden, in half of the BUILD primary sites, we sampled students in the biomedical fields, with a target of collecting over 400 biomedical students' responses per site per survey. As a result, in the sampled sites, the majority of the survey participants were in the biomedical and STEM-related fields. In the other half of the BUILD primary sites, since the student populations were relatively small, we sent the surveys to the entire continuing undergraduate cohort. Around a quarter of those who had a TFS response record in the previous year also responded to the SAFS. For the BUILD scholars, due to the program incentive, nearly 80% of the TFS participants responded to the SAFS during the spring of their freshman year.

Although the first BUILD scholar cohort was the 2015 cohort, we only include cohorts of 2016 to 2019 in our analysis, because the TFS 2015 did not provide a baseline assessment of the outcome, and the first SAFS administration occurred in the spring of 2017. Typically, incoming freshmen completed the TFS before the fall of their first year and the SAFS during the spring of their first year. For example, a student who entered college the fall of 2016 would have completed the TFS 2016 before the start of the 2016 fall semester and then completed the SAFS 2017 in the spring of 2017. Out of the 205 first-year BUILD scholars who had a record in the TFS, 162 responded to the SAFS in the following spring, and 134 students completed the outcome variable-related question in the SAFS questionnaire. All of the 134 students graduated from high school during the same year they entered into college, and their first fall semester in the BUILD sites was the first time they enrolled into college as undergraduate students. All first-year BUILD scholars were full-time students, and planned to obtain a bachelor degree or above. By the time they responded to the SAFS in the spring of their first year, most of them had received several months of BUILD scholar interventions, although the amount of interventions they were exposed to might vary case by case due to the long survey response windows.

We selected students in the 4 studied sites who completed the TFS and the following SAFS with no missing responses in key variables that we were interested in, and who were not in the BUILD scholar program in their first year, into the control group. To mimic the BUILD scholar group sample, we limited the control group students to full-time first-time college students who were straight out of high school when they started college, and who planned to obtain a bachelor degree or above. We ended up with 1988 students in the control group, and combined with 134 BUILD scholars, our initial sample size was 2122.

4.4.2 Variable Construction

Following the previous framework (Figure 4.3), there were 3 types of variables in this study — the outcome variable (Y) that measured students' intent to pursue science-related careers, the intervention indicator (D), and covariates (X) that potentially associated with Y and/or D (If a covariate X solely associates with D, and D impacts Y, then D is a mediator in the causal chain). We summarized the variables and their coding in Table 4.4. In this table, we also marked the reference group for the categorical variables in regression analyses. In the next few paragraphs, We introduced how we mapped the variables from survey items, and explain the variable coding (Table 4.4) in detail.

Table 4.4: Variable List

	Variable	Name	Coding
Dependent Variable	Science Career Intent	Science Career	Definitely no (1), Possibly no (2), Uncertain (3), Possibly yes (4), Definitely yes (5)
Intervention	BUILD Scholar Designation	Scholar	Scholar, Control group student*
Covariates	Race/Ethnicity	Race	Asian, Black, Hispanic, White*, Other, and Two or more race/ethnicity
	Gender Identity	Gender	Male*, Female, Others
	First Generation College Student Status	First-gen	First-gen, Non-first-gen*
	Pell Grant Status	Pell	Pell Grant receiver, Did not receive Pell Grant*
	High School GPA	High School GPA	A or A+ (8), A- (7), B+ (6), B (5), B- (4), C+ (3), C (2), D (1)
	Years of Math prior to College	Math Training	None (1), 1/2 (2), 1 (3), 2 (4), 3 (5), 4 (6), 5 or more (7)
	Science Identity	Science Identity	EAP scores, mean = 5, variance = 1
	Researcher Self-Efficacy	Researcher Self-Efficacy	EAP scores, mean = 5, variance = 1
	Baseline Science Career Intent	Baseline	Definitely no (1), Possibly no (2), Uncertain (3), Possibly yes (4), Definitely yes (5)
	Degree Expectation	Degree	Bachelor*, Graduate, Biomedical and academic terminal
	Site	Site	A, B, C, D*
	Cohort	Cohort	2016*, 2017, 2018, 2019
	Major	Major	Non-biomedical*, Biomedical social science, Biomedical natural science

*Reference group in regression models.

4.4.2.1 Outcome Variable

The outcome variable or dependent variable, students' intent to pursue science-related careers, was measured by a survey item in the SAFS: Will you pursue a science-related research career? The response categories were coded from 1 to 5, representing "Definitely no," "Possibly no," "Uncertain," "Possibly yes," and "Definitely yes." This question was also asked in the TFS, of which the responses were used as a baseline measure of the outcome.

4.4.2.2 Intervention Indicator

The intervention indicator was a dichotomous variable linked from the program participation data to the survey data that suggested if a students was in the treatment group (for scholars, D = 1) or control group (for non-scholars, D = 0) between the time they took the TFS and SAFS. Students identified as in the treatment group were BUILD scholars who were first year students selected through the criteria in Table 4.3, and participated the activities in Table 4.2, between the time period that they responded to the TFS and corresponding SAFS.

4.4.2.3 Covariates

We identified the covariates based on the previously mentioned influential factors in the literature, and summarized them into the suggested 4 categories: demographic variables, pre-college academic performance and experience in STEM education, self-efficacy in science and research, and college experience. All of the covariates were extracted and computed from the responses to the TFS.

Demographic variables. The demographic variables included race/ethnicity, gender identity, first generation college student status, and Pell Grant status. The following sections provide further details about how each of the models operationalized each of these characteristics.

Race/Ethnicity. In the surveys, we asked students to mark all that apply: White/ Caucasian, African American/Black, American Indian/Alaska Native, East Asian (e.g., Chinese, Japanese, Korean, Taiwanese), Filipino, Southeast Asian (e.g., Cambodian, Vietnamese, Hmong), South Asian (e.g., Indian, Pakistani, Nepalese, Sri Lankan), Other Asian, Native Hawaiian/Pacific Islander, Mexican American/Chicano, Puerto Rican, Other Latino, and Other. We then regroup students into Race/Ethnicity Groups: American Indian, Asian, Black, Hispanic, White, Other, and Two or more race/ethnicity. In this study, we combined the American Indian group into the Other group, due to a lack of observations of that category in the BUILD scholar group.

Gender. We had different versions of asking students' gender identity. In general, we grouped those who self-identified as Male, Man or Trans Man as Male, Female, Woman or Trans Women as Female, and others such as Gender queer, Gender non-conforming or Different identity as Others.

First generation college student status. The first generation college student status, or first-gen status, was a computed variable from students' parents' education. In ours study, we used the definition of first generation college students provided in Maccalla et al. (2020), and assigned value "1" for those students whose parent(s) did not have a bachelor degree, and value "0" for students with at least one parent with a bachelor degree or above. We made this decision also because our study sites were all 4-year colleges, and participants were full-time college students. It would be an advantage for students whose parent(s) successfully went through college education under relatively equivalent settings.

Pell Grant status. We used Pell Grant receiving status to reflect students' socioeconomic status. Students who received Pell Grant were coded as "1" and those who did not as "0" on this variable.

Pre-college academic performance and experience in STEM education. The related variables are high school GPA and math training prior to college. We did not include students' SAT or ACT performance because not all sites specified them as an admission requirement.

High school GPA. We asked students to respond to the question, "What was your average grade in high school? (Mark one)" and provided the options: A or A+, A-, B+, B, B-, C+, C, and D, coded as numbers from 8 to 1.

Math training prior to college. For math training prior to college, we focused on students' years of math training in high school, and used the responses to the question

“During high school (grades 9-12) how many years did you study each of the following subjects?” on the subject Mathematics, and the options were coded from 1 to 7, representing: None, 1/2, 1, 2, 3, 4, and 5 or more.

Self-efficacy in science and research. In addition to science identity and researcher self-efficacy, in this special case that our outcome variable also reflected self-efficacy in science and research, the baseline measure of students’ intent to pursue science-related careers, and similarly, students’ degree expectation (another important outcome to study in the future) were also included as a covariate.

Science identity. Since we used the definition of science identity from the HERI, we also used the scale developed by the HERI (CIRP Constructs, n.d.). The scale asked students’ to indicate their level of agreement (coded from 5 to 1, Strongly Agree to Strongly Disagree, with 3 being the Neutral option on the 4 questions) with the following statements: “I have a strong sense of belonging to a community of scientists,” “I derive great personal satisfaction from working on a team that is doing important research,” “I think of myself as a scientist,” and “I feel like I belong in the field of science.” The variable was quantified using students’ expected-a-posterior (EAP) item response scores on the HERI national sample in 2016, and was originally centered and scaled at N(50, 10). We rescaled it to N(5, 1) for the convenience of interpretation in future analyses.

Researcher self-efficacy. Similarly, we used the 10-item HERI scientific researcher self-efficacy scale to measure students’ researcher self-efficacy (CIRP Constructs, n.d.). Students were asked to rate their confidence level from 5 to 1, representing Absolutely, Very, Moderately, Somewhat, and Not at All on the 10 questions: “Use technical science skills (use of tools, instruments, and/or techniques),” “Generate an answerable research question,” “Determine how to collect appropriate data,” “Explain the results of a study,” “Use scientific literature to guide research,” “Integrate results from multiple studies,” “Ask relevant questions,” “Identify what is known and not known about a problem,” “Understand scientific concepts,” and “See connections between different areas of science and mathematics.” The HERI conducted the scoring (EAP score) on the

2016 national sample, and centered the scores to N(50, 10). Again, we rescaled it to N(5, 1) for the convenience of interpretation in future analyses.

Baseline of students' intent to pursue science-related careers. As mentioned previously, this variable was from students' responses to "Will you pursue a science-related research career?" in the TFS.

Degree expectation. We asked the question "What is the highest academic degree that you intend to obtain?" and provided the options: None, Vocational certificate, Associate (A.A. or equivalent), Bachelor's (B.A., B.S., B.D., etc.), Master's (M.A., M.S., M.B..A., etc.), J.D. (Law), M.D., D.D.S., D.V.M., etc. (Medical), Ph.D., Professional Doctorate (Ed.D., Psy.D., etc.), and Other. We categorized the degree options "None, Vocational certificate, Associate (A.A. or equivalent)" as Below bachelor, "Bachelor's (B.A., B.S., B.D., etc.)" as Bachelor, "Master's (M.A., M.S., M.B..A., etc.), J.D. (Law), Professional Doctorate (Ed.D., Psy.D., etc.)" as Graduate, and "M.D., D.D.S., D.V.M., etc. (Medical), Ph.D." as Biomedical and academic terminal. As mentioned earlier, all our participants planned to obtain at least a bachelor degree or above, so the categories of this variable that appeared in this study were only Bachelor, Graduate, and Biomedical and academic terminal.

College experience. Our participants were from 4 different BUILD sites across 4 different cohorts, which made it extremely challenging to identify well-defined measurable common college experience in this study. Therefore, instead of identifying specific college experience, we identified covariates that might bring different types of college experience. We believed that students from different institutions, different cohorts and different majors could have quite different experiences in college.

Site. The 4 sites in this study were coded as site A, B, C, and D, due to the identification protection purpose, for some BUILD scholar programs only enrolled less than 5 students of a certain class standing in a cohort. The names of the sites could be reviewed internally for program improvement purposes.

Cohort. The cohort variable was coded as 2016, 2017, 2018, and 2019, indicating which year the participants started college.

Major. The majors were originally reported by students in the TFS, and we coded them into three categories: Non-biomedical, Biomedical social science, and Biomedical natural science.

4.4.3 Data Analysis

Under the framework presented in Figure 4.3, we assumed that the covariates confounded both Y and D. We proposed two different approaches that emphasized different perspectives. In the first approach, we applied a matching strategy that is similar to the 2SM to ensure the local similarity at the largest extent. After examining the covariance balance, we used regression to estimate program effects. This approach was suitable under the conditions that the covariance balance was off and we had many observations in the control group. In the second approach, we applied the regression analysis on the whole dataset and then performed sensitivity analysis to examine the plausibility of the program effects and the influence of potential confounders. This approach might be useful when we have a small number of observations in the dataset or the covariates were relatively balanced.

4.4.3.1 Approach One: Multi-stage Matching

Based on the selection criteria (Table 4.3), we employed multi-stage matching to address potential selection bias attributable to the fact that program participation was not randomly assigned to students at the BUILD sites. The philosophy in this approach was to use matching to identify students in the control group who were similar to the BUILD scholars in this multisite setting, to achieve conditional ignorability. After assessing the covariance balance of the matched data, we used the matched data to perform regression analysis and estimating the program effects.

The multi-stage matching procedure, revised from the 2SM developed by Rickles and Seltzer (2014), was designed uniquely for the features of the BUILD programs, which could be used for matching data from programs that implemented interventions in multiple institutions and/or across multiple cohorts (and/or other group level indicators), especially when school and/or cohort (and/or other group level indicators) level sample size was less than 20. This matching procedure accounted for the heterogeneity of the across group level (e.g., between-site) treatment effects, and it was superior to matching at single level separately (Rickles & Seltzer, 2014) because it allowed us to borrow information from participants in other sites. This would be especially helpful in our case, since some single-site/cohort might have a small sample size that limited the statistical power as well as limited the likelihood of finding a close match.

The number of stages depended on if and when the number and the quality of the matched cases met the researchers' tolerance, as Ho, Imai, King, and Stuart (2007) described that matching was a nonparametric process. In this procedure, identifying local matches was our priority. We used the matching on the BUILD scholar program participation as an example to demonstrate each step. Out of the 2122 observations in the data, 134 students were BUILD scholars. After many attempts, we proposed the following procedure.

Stage one: identify within-group matches. We assumed that the across group level treatment effect heterogeneity of the BUILD scholar program existed at both the institutional level and the cohort level. This assumption was based on the fact that each site had their own selection criteria and the competition of getting into the scholar program might vary across different cohorts. Therefore, we considered the cross-classification of site and cohort as the group level in this study. Table 4.3 indicated that sites would select their BUILD scholars based on (apart from first-item enrollment) high school GPA, major, degree expectation, baseline intent to pursue science-related research careers, science identity, and research self-efficacy. We used a logistic regression (Eq. 4.1) to estimate the propensity for BUILD scholar program participation, which was notated as $P(D_{ijk} = 1)$ for student i in site j and cohort k.

$$\text{logit}(P(D_{ijk} = 1)) = \boldsymbol{\beta_{jk}} \cdot \left[1, site_{ij}\mathbf{X_{sc}}, cohort_{ik}\mathbf{X_{sc}}, site_{ij} \cdot cohort_{ik} \cdot \mathbf{X_{sc}}, \mathbf{X_g}\right]^\top \tag{4.1}$$

In Eq. 4.1, the coefficients $\boldsymbol{\beta_{jk}} = (\beta_{0jk}, \beta_{1jk}, \ldots, \beta_{pjk})$, where p is the number of covariates for estimating the propensity of BUILD scholar program participation for students in site j and cohort k. The indicators, $site_{ij}$ and $cohort_{ik}$, represented students' memberships of BUILD sites and cohorts. We used $\mathbf{X_{sc}}$ to represent the covariates whose association with BUILD scholar program participation might be strongly influenced by site and cohort memberships, and $\mathbf{X_g}$ to represent the rest of covariates that generically influenced BUILD program participation across groups.

We did not use a multilevel structure to estimate the propensity scores, because the group level ($site \times cohort$) sample size was small ($n_{group} = 4 \times 4 = 16$); instead, we used interactions of the group membership and selection criterion related variables to create unique group level estimates. We considered other covariates as universally influenced students' overall chance of being selected into a diversity training program in all sites. We, as suggested by Ho et al. (2007), used the logistic regression with logit link (Eq. 4.1) to predict the propensity scores and identified the 2 nearest neighbors (to keep more observations and to maximize the analytical sample size), defined by the Mahalanobis distance, for each BUILD scholar using potential large-effect variables: all variables in $\mathbf{X_{ijk}}$, with the continuous variables being mean-centered.

We set a caliper of .25 standard deviations (Cochran & Rubin, 1973) of the propensity scores estimated using all covariates excluding site and cohort, to ensure that the propensity scores of the matched control units were within .25 standard deviations from that of the treated unit they matched to. Although .25 was a common caliper in matching, we could adjust the caliper based on our preference of the amount of local matches, or in other words, on our prior knowledge about the program differences at the group level (Rickles & Seltzer, 2014). At this stage, we only allowed treated units to be matched, without replacement, to control units within the same group (exact matching on site and cohort) — e.g., a BUILD scholar of cohort 2016 from site A should be matched with two other non-BUILD scholars from site A who entered site A as freshmen in 2016.

We might end up with two matched control units for each treated unit after stage one. If so, we could move on to the regression analysis. If not, which might be more likely, we would move on to the next stage. The illustration could be similar to the Figure 1 in Rickles and Seltzer (2014, p. 619).

Stage two: identify within-site matches. After stage one, we might end up with a certain percentage of BUILD scholars that found two matched control cases, some only found one matched control unit and some found none. At stage two, we relaxed the constraints to allow cross cohort matching for those treatment units that failed to find two matched control units from the previous matching. We still only allow the treatment units to be matched with control units in the same site, but could be matched with control units in all other cohorts apart from their own (exact matching on site). For example, a BUILD scholar in site A cohort 2016 can only be matched with students unaffiliated with the BUILD program in site A cohort 2017, 2018 or 2019. We used the same matching standard, in which we used Mahalanobis distance matching on the same set of selected covariates as in the previous step, with a caliper of .25 of the propensity scores computed from all covariates.

The only complication for this stage was that, in the previous matching, the distances and the propensity scores were “group-based” and therefore, not comparable across different site and cohort combinations (Rickles & Seltzer, 2014). To address this issue, we recalculated the distances and propensity scores using the respective subset of variance-covariance matrix and parameter estimates of the treatment units that the control units matched to. For example, a student in the control group, who might come from site A cohort 2017, 2018 or 2019, was about to be matched with a BUILD scholar in site A cohort 2016, we would use the subset of variance-covariance matrix and parameter estimates for the group site A cohort 2016 to rescale the control units’ distances and propensity score, “pretending” that this student was from the same group as the treatment unit that this control unit was about to matched to. This simple transformation ensured that the across cohort matching was based on the same propensity score model (Rickles & Seltzer, 2014).

We would skip those treatment units that already had two matched control cases from the previous matching. For those BUILD scholars who only found one local matched control unit at stage one, we selected the "best fit" of the matched control cases defined by the minimal Mahalanobis distance, within the caliper. For the rest of the BUILD scholars who failed to find any matched control units, we selected the top two fits of the matched control cases within the caliper. If the distances of multiple matched cases were the same (very unlikely, almost impossible), we would choose the one from the nearest cohort. As in the previous stage, we would stop the matching process if we could identify two matched control units for all treated units.

Stage three: identify cross-site matches. Following the previous matching results, we relaxed the constraints to allow cross-site matches. Similar to previous stages, we use distance matching on large-effect covariates. We planned to find two matched control units for all treated units, and to meet this goal, at this stage, we did not set a caliper. We allow the treatment units to be matched with control units from different sites other than their own sites. To ensure the comparability of the distances, we would use the subset of variance-covariance matrix of the treatment unit's site of belonging to rescale the control units' distances.

Balance assessment. After matching, we assessed the covariate balance through comparing the raw (unweighted) means of covariates between the BUILD scholars and the matched students in the control group. We also presented the standardized mean difference (a measure of effect size), variance ratio (VR) and Kolmogorov–Smirnov (KS) statistics to assess the differences of the covariates' empirical distributions between the treated and the control groups.

Outcome adjustments. We adjusted the outcome variable for those control units that matched across cohorts and/or sites, using the strategy proposed by Rickles and

Seltzer (2014). The general idea was to use the group level mean differences to construct the outcome values, as if the matched control units were from the same group as the treatment units they were matched to. We used site and cohort associated estimates from Eq. 4.1 to perform adjustments following Eq. 4.2.

$$\begin{gathered} Y^*(0)_{ijk} = Y(0)_{ij'k'} + \mathbf{X_i} \cdot (\boldsymbol{\Delta}\mathbf{site} + \boldsymbol{\Delta}\mathbf{cohort} + \boldsymbol{\Delta}\mathbf{sc}) \\ \boldsymbol{\Delta}\mathbf{site} = \mathbf{b_j} - \mathbf{b_{j'}},\ \boldsymbol{\Delta}\mathbf{cohort} = \mathbf{b_k} - \mathbf{b_{k'}},\ \boldsymbol{\Delta}\mathbf{sc} = \mathbf{b_{jk}} - \mathbf{b_{j'k'}} \end{gathered} \tag{4.2}$$

In Eq. 4.2, $Y^*(0)_{ijk}$ was the adjusted outcome value for a control group student i who matched to a BUILD scholar in site j and cohort k. $Y^{(}0)_{ijk}$ was the original outcome value for this control group student i who was originally from site j' and cohort k'. We used **Δsite**, **Δcohort** and **Δsc** to represent the site level, cohort level and site cross cohort level differences, respectively. To compute these differences, we used the group associated estimates from Eq. 4.1. For example, **Δsite** was computed from $\mathbf{b_j} - \mathbf{b_{j'}} = (b_{0j} - b_{0j'}, b_{1j} - b_{1j'}, \ldots, b_{qj} - b_{qj'})^\top$, where $b_{\cdot j}$ represented site related covariate estimates for site j, $b_{\cdot j'}$ represented site related covariate estimates for site j', and q was the number of covariates. $\mathbf{X_i}$ was a $m \times (p+1)$ design matrix with the first column to be constant 1 and the rest of columns each representing students' responses to a variable in $\mathbf{X_{sc}}$, and m was the number of students who were cross-matched in stage two and stage three.

Regression analysis. With the matched sample, we ran OLS regression on the 402 observations (134 BUILD scholars and 268 non-BUILD students) and controlled for covariates that were used in matching to estimate the BUILD scholar program impact on students' intent to pursue science-related research careers. Although for the purpose of estimating program effects, we did not have to include the covariates since we could possibly rule out the correlation between the intervention and the covariates through the matching process, we included the covariates to help us understand what influenced students' intent to pursue science-related research careers in general.

4.4.3.2 Approach Two: Sensitivity Analysis

In addition to using strategies to ensure conditional ignorability, we could use sensitivity analysis to conduct a post-hoc assessment of the confounder issue. This approach might be suitable for the situation that the initial balance of the covariates were good, or the number of observations were so limited that researchers did not want to discard any case. This approach was less complicated than the previous one, but the interpretation of the results required the familiarity of the context related literature. The advantage of using sensitivity analysis was that it not only informed the potential impact of observed confounders included in the analysis, but also provided benchmarks as references for assessing the possibility of the existence of unobserved confounders. In addition, if the sensitivity analysis indicated that the impact of the treatment effect on the outcome variable was not jeopardized by confounders, the regression results could be more likely to be generalized to a broader population.

The major procedure included two steps. We firstly conducted regression analysis using the full observed data, and then performed sensitivity analysis with benchmarks (Cinelli & Hazlett, 2020) that were strategically selected based on the regression results. In this study, we used the dataset to demonstrate this approach as a convenient example to present the methodological differences, regardless of the suitability of applying this approach to analyze the data.

Regression analysis. In this analysis, we used all 2122 observations in the dataset, including 134 BUILD scholars and 1988 non-BUILD students. We ran OLS regression on the treatment indicator, controlling for all covariates that were mentioned in the previous sections to estimate the BUILD scholar program impact on students' intent to pursue science-related research careers.

Sensitivity Analysis. We then ran a sensitivity analysis based on the results from the OLS regression to examine if the observed covariates confounded with the treatment and the outcome variables and the possibility of having unobserved confounders that largely influenced the conclusion of the program effects. Unobserved confounders are variables that were not included in the analytical model but might influence the results. What if, hypothetically, we left out a variable that was really important to the model and might have changed the results? We used sensitivity analysis to evaluate if the regression results were sensitive to observed covariates and unobserved confounders or to what extent the results might have changed due to observed covariates or unobserved confounders.

4.5 Results

In this section, we reported the analytical results in the order of the two approaches described in the previous section. Along with reporting the results, we focused on the interpretation of the results and addressed the research question, "Does participation in the BUILD scholar program during freshman year impact students' intent to pursue science-related research careers?"

4.5.1 Approach One: Multi-stage Matching

4.5.1.1 Matching

We chose pre-intervention variables and background demographic variables as covariates in the matching process, because from the previous review of literature and the program, we believed that those variables might influence the likelihood of being selected into the BUILD scholar program. That included the students' self-selection of applying for the program, as well as the sites selecting students based on their selection criteria.

Table 4.5 presented the covariate mean comparisons of pre- and post-matching between the BUILD scholar group and the control group. We used the standardized mean difference (SMD), a measure of effect size, to assess the magnitude of the mean differences. For continuous variables, such as high school GPA, math training, science

Table 4.5: Comparisons of Covariate Means and SMDs

	Scholars (mean)	Control Group (mean)		SMD	
		Pre-match	Post-match	Pre-match	Post-match
Sex: Male	0.306	0.338	0.279	-0.032	0.027
Sex: Female	0.672	0.638	0.691	0.033	-0.020
Sex: Others	0.022	0.024	0.029	-0.001	-0.007
Pell	0.433	0.349	0.386	0.084	0.047
First-gen	0.284	0.281	0.257	0.003	0.026
High School GPA	6.948	6.807	7.040	0.118	-0.082
Baseline: Definitely no	0.097	0.124	0.074	-0.027	0.024
Baseline: Possibly no	0.075	0.187	0.081	-0.113	-0.006
Baseline: Uncertain	0.134	0.221	0.129	-0.087	0.006
Baseline: Possibly yes	0.224	0.232	0.221	-0.009	0.003
Baseline: Definitely yes	0.470	0.235	0.496	0.235	-0.026
Major: Non-biomed	0.022	0.257	0.018	-0.235	0.004
Major: BM Social	0.082	0.066	0.033	0.016	0.049
Major: BM Natural	0.896	0.677	0.949	0.219	-0.053
Race: White	0.194	0.385	0.239	-0.191	-0.045
Race: Asian	0.187	0.178	0.250	0.009	-0.063
Race: Black	0.284	0.104	0.210	0.180	0.074
Race: Hispanic	0.105	0.165	0.107	-0.060	-0.002
Race: Other	0.030	0.016	0.022	0.014	0.008
Race: Two or more	0.202	0.152	0.173	0.049	0.029
Degree: Bachelor	0.022	0.193	0.011	-0.170	0.011
Degree: Graduate	0.269	0.443	0.228	-0.174	0.041
Degree: Terminal	0.709	0.365	0.761	0.344	-0.052
Math training	5.963	5.953	5.996	0.020	-0.066
Science identity	6.153	5.414	6.102	0.934	0.077
Research self-efficacy	5.506	5.290	5.448	0.237	0.071

identity and research self-efficacy, we compared their mean differences. For other variables that coded categorically, we compared the proportions of categories that were observed in the treatment and the control groups. We coded the baseline students' intent to pursue science-related careers as categorical in the matching procedure, because we wanted to ensure that the number of observed categories of the baseline to be proportionately similar.

Table 4.6: Comparisons of Variance Ratio (VR) and Kolmogorov–Smirnov (KS) Statistics

	Pre-matching		Post-matching	
	VR	KS Statistics	VR	KS Statistics
High School GPA	0.838	0.050	1.188	0.030
Math Training	1.089	0.009	1.319	0.031
Science Identity	0.528	0.376	0.980	0.098
Research Self-efficacy	0.675	0.152	0.868	0.077

In the pre-matching sample, we observed some covariate imbalance between the treatment group and the control group. BUILD scholars had higher high school GPA ($SMD = .118$), higher science identity ($SMD = .934$) and higher research self-efficacy ($SMD = .237$) than the control group students. For the continuous covariates, only the variable math training had an absolute SMD that was smaller than 0.1, a threshold (Stuart, Lee, & Leacy, 2013) for determining the achievement of satisfactory balance (the SMDs that closer to 0, the better). For continuous variables, we also examined the VR and the KS statistics of the BUILD scholars and the control group students (Table 4.6). We observed that the VRs of the science identity ($VR = .528$) and research self-efficacy ($VR = .675$) were far from 1, the VR benchmark that indicated the equal variance in the two groups. The two-sample KS statistics, which measured the largest distance of the empirical cumulative density functions (eCDFs) between the two sample, of the science identity ($KS = .376^{***}, p < .001$) and research self-efficacy ($KS = .152^{**}, p < .01$) were significant and a lot larger than 0, the value that implied perfectly identical distributions. The VRs and KS statistics further confirmed the imbalance of science identity and research self-efficacy between the two groups.

Among covariates that were categorical (including dichotomous) variables, gender identity, Pell Grant status, first generation college student status were balanced. There was a higher percentage of BUILD scholars who selected "5" in the baseline question ($SMD = .235$), meaning that proportionately, more BUILD scholars definitely intended to pursue science-related research careers before they started college. Similarly, a higher percentage of BUILD students chose (broadly defined) biomedical natural science majors ($SMD = .219$), and a higher percentage of BUILD scholars planned to obtain biomedical and academic terminal degrees ($SMD = .344$). Compared to the control group students, a higher proportion of BUILD students were self-identified as Black/African American ($SMD = .180$), and a lower percentage of BUILD students were self-identified as White ($SMD = -.191$).

We expected these differences, because they reflected the selection criteria of this undergraduate diversity training program. In general, the BUILD scholars had stronger self-efficacy in science and research at the baseline (before they started college), and were more diverse in race/ethnicity than the control group students. The observed covariate imbalance suggested that some of the covariates confounded the treatment and the outcome, which indicated that strategies such as matching were probably necessary to help address the selection bias.

We strictly followed the procedures described previously, and identified two matched control cases for each BUILD scholar. At the first stage, we were able to identify 206 matched control units for 114 BUILD scholars, including 92 scholars who were matched with 2 control units, 22 scholars with 1 control unit, and 20 scholars who were not matched with any control unit. We moved on to stage two, hoping to find matched cases for the 42 scholars ($134 - 92 = 42$). Stage two was not very productive. We only helped 12 scholars find one matched case, out of the 22 scholars who found a match in stage one; in addition, we helped another 7 scholars find one matched case, out of the 20 scholars who found no match previously. After stage two, in total, 104 scholars successfully found two matched cases, 17 scholars had one matched case, and 13 scholars had none. We moved forward to stage three, and found matched cases for the 17 scholars who had one matched case and 13 scholars who had none in the previous stages. Through the whole process, we matched the cases without replacement. We found 76.87% control units at stage one, 7.09% units at stage two, and 16.04% units at stage three. If we were about to redesign the procedure, we would probably consider dropping the caliper at stage two to finish matching there, or skipping stage two and directly using the matching criteria in stage three, since stage one captured over three quarters of the matched control cases and the cohort differences seemed to be subtle.

After matching, the covariate balance was improved for almost all covariates. All post-matching absolute SMDs were below .1 (Table 4.5). We plotted the absolute SMD comparisons of pre- and post-matching samples in Figure 4.4. The dash line marked the .1 threshold that we used for assessing satisfactory balance. The VRs and KS statistics (Table 4.6) also indicated balance improvement for previously problematic covariates: science identity and research self-efficacy, since the post-matching VRs were closer to 1 and the KS statistics were no longer significant. In addition, although the site and cohort were not included as covariates for matching distance computing, the site variable for students in the two groups was exact at stage one and two, and the cohort variable was exact at stage one. The proportions of site of belong and cohort of belonging were relatively balanced across the two groups (Table 4.7). After matching, the covariate balance improved, and we were more confident about meeting conditional ignorability with the matched data.

Before we used the matched data to run OLS regression, we adjusted the outcome variable for those control group units that matched to treatment units in different cohorts and/or sites. We performed the transformation based on the Eq. 4.2 for the control units matched at stage two and stage three. For those matched at stage two, the adjustment would only account for the cohort differences, meaning that $\boldsymbol{\Delta}\mathbf{site} = 0$ and

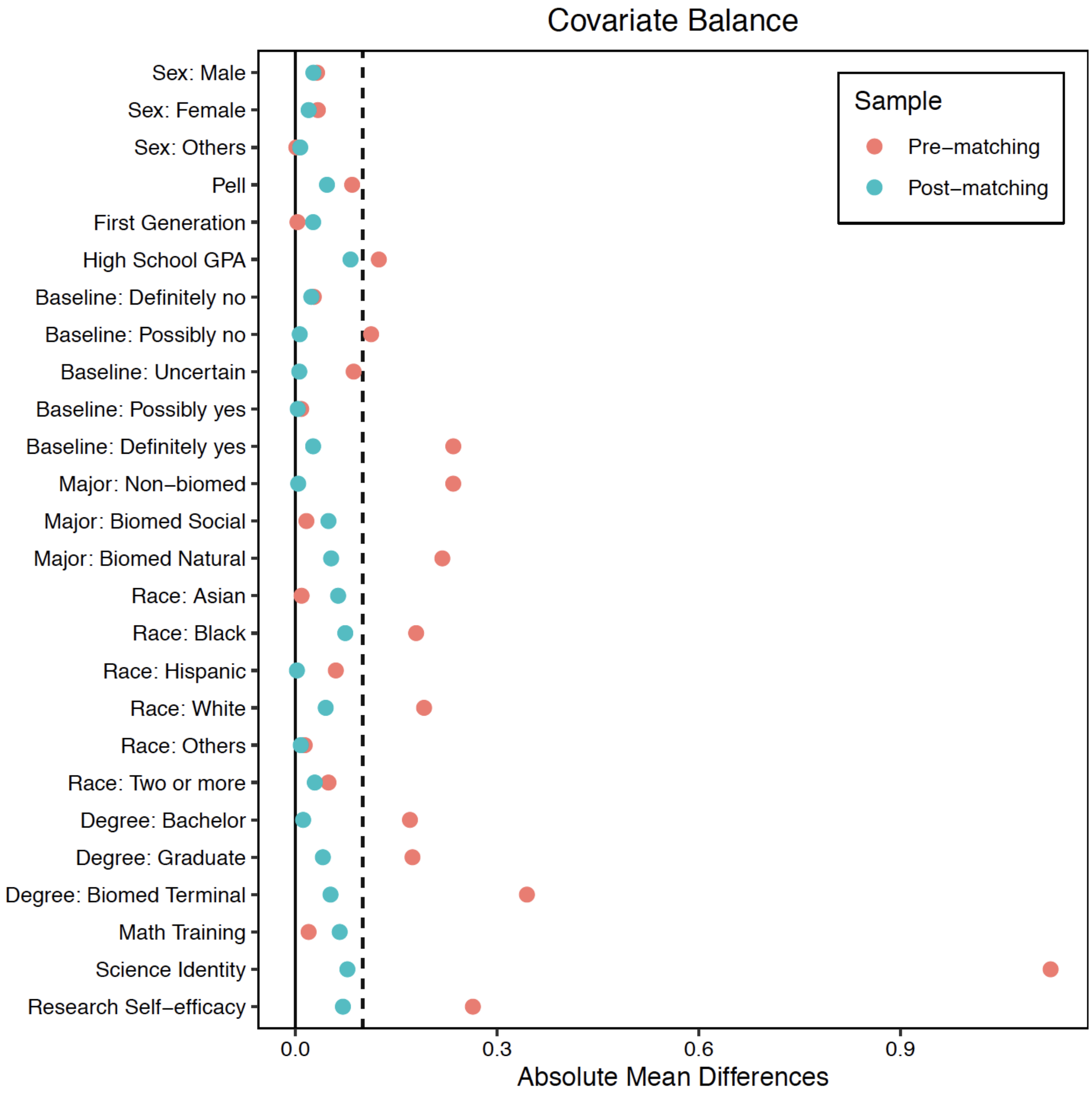


Figure 4.4: Covariate Balance: Absolute Standardized Mean Differences

Note. The Standardized Mean Difference values were presented in Table 4.5.

Δsc = 0. The mean of the original measure of the outcome variable of the 62 matched control units that needed the transformation was 4.129, and after the adjustment, it decreased to 4.118. The variance changed from 1.236 to 1.229. This result indicated that the adjustment process might not be necessary, considering it only affected less than a quar-

Table 4.7: Comparisons of Site and Cohort Balance

	BUILD Scholars	Control Group	
		Pre-matching	Post-matching
Site: A	5.2%	11.8%	6.3%
Site: B	38.8%	25.8%	26.5%
Site: C	46.3%	38.3%	53.3%
Site: D	9.7%	24.1%	14.0%
Cohort: 2016	29.9%	16.6%	26.5%
Cohort: 2017	18.7%	18.3%	15.1%
Cohort: 2018	33.6%	37.4%	38.6%
Cohort: 2019	17.9%	27.7%	19.9%

ter of observations and from a post-hoc perspective, the mean and variance differences between the original outcome values and the adjusted values were subtle. Nevertheless, we inserted the adjusted outcome values of the 62 students as their post-measure values of students' intent to pursue science-related research careers in the regression analysis.

4.5.1.2 Regression (Sample Size = 402)

As mentioned, to estimate program effects, under the condition that the covariates were balanced between the two groups, a simple regression that regressed the outcome variable (Y) on the intervention variable (D) would help us understand the impact of the program. In the regression analysis, we treated both the baseline and post-measure of students' intent to pursue science-related research careers as continuous variables, since they were ordinal in nature (rating from 1 to 5, representing "definitely no" to "definitely yes" to pursue science-related research careers). Table 4.8 presented the analytical results of this simple regression model, which majorly showed that on average, the BUILD scholars' post-measure of intent to pursue science-related research was .267 ($p = .004$) higher than their peers in the control group. This result was significant at $p < .01$ level, and the treatment explained 1.7% of the post-measure outcome. In this model, the BUILD scholar program effect size (Hedges' g, recommended by the What Works Clearinghouse, 2021) was 0.275. This model helped us answer the research question that participation in the BUILD scholar program during freshman year positively impacted students' intent to pursue science-related research careers.

Table 4.8: Simple Regression Model (N = 402)

	Estimate	Std. Error	p value	sig.level
(Intercept)	4.053	0.061	0.000	***
Scholar	0.267	0.093	0.004	**

Adjusted R-squared: 0.017

BUILD scholar program effect size (Hedges' g): 0.275 (SE = 0.106)

sig.level: $^{***}p < 0.001,^{**} p < 0.01,^{*} p < 0.05$

In addition to including the intervention variable in the model, we were also interested in including the covariates to help us understand what influenced students' intent to pursue science-related research careers in general. Table 4.9 presented the results of the model that included the covariates used in the matching process. In this model, our conclusion regarding the BUILD scholar program effects on students' intent to pursue science-related research careers stayed the same. On average, the BUILD scholars' post-measure of intent to pursue science-related research was .265 ($p = .004$) higher than their peers in the control group and the estimate was significant at $p < .01$ level. The BUILD scholar program effect size was 0.274. Judging from the similarity of the treatment estimates in the two models, the covariates did not influence the treatment effects, indicating the likelihood of confounding was low.

This model explained 28.5% of the outcome variable. We observed that students' baseline measure of intent to pursue science-related research careers, science identity and degree expectations were the most significant covariates, after controlling for other covariates and the intervention effects. Students who rated 1 unit higher on the baseline

Table 4.9: Multiple Regression Model (N = 402)

	Estimate	Std. Error	p value	sig.level
(Intercept)	0.088	0.812	0.914	
Scholar	0.265	0.091	0.004	**
Site: A	-0.025	0.251	0.922	
Site: B	-0.024	0.17	0.889	
Site: C	0.025	0.174	0.885	
Cohort: 2017	0.074	0.151	0.626	
Cohort: 2018	0.007	0.107	0.949	
Cohort: 2019	0.087	0.132	0.509	
Sex: Female	-0.009	0.097	0.925	
Sex: Others	0.099	0.238	0.678	
Pell	-0.026	0.097	0.788	
First-gen	-0.019	0.104	0.856	
High School GPA	0.085	0.042	0.046	*
Baseline	0.262	0.039	0.000	***
Major: Biomed Natural	0.793	0.315	0.012	*
Major: Biomed Social	0.653	0.361	0.071	
Race: Asian	-0.052	0.129	0.688	
Race: Black	0.055	0.132	0.679	
Race: Hispanic	0.081	0.187	0.668	
Race: Other	-0.586	0.362	0.107	
Race: Two or more	0.051	0.147	0.729	
Degree: Academic Terminal	1.168	0.348	0.001	***
Degree: Graduate	0.901	0.346	0.010	**
Math Training	-0.155	0.092	0.092	
Science Identity	0.278	0.078	0.000	***
Research Self-efficacy	-0.051	0.057	0.364	

Adjusted R-squared: 0.285

BUILD scholar program effect size (Hedges’ g): 0.274 (SE = 0.106)

sig.level: $p < 0.001^{***}, p < 0.01^{**}, p < 0.05^{*}$

measure of students' intent to pursue science-related research careers, rated .262 ($p <$ .001) unit higher on the outcome variable. On average, one standard deviation higher on the science identity led to .278 ($p <$ .001) unit higher on the outcome variable. Compared to students who did not plan to obtain a graduate degree, those who planned to obtain a graduate degree and who planned to obtain a biomedical and academic terminal degree leveraged their intent to pursue science-related research careers by .901 ($p <$ .01) and 1.168 ($p <$ .001) units, respectively. Students who were in biomedical natural science majors rated .793 ($p <$ .05) higher on the outcome variable. Students who had higher high school GPA were more likely to intend to pursue science-related research careers ($p <$ .05), although the magnitude was rather small.

4.5.2 Approach Two: Sensitivity Analysis

In the second approach, we used 2122 observations in the dataset, including 134 BUILD scholars and 1988 non-BUILD students. Although we found confounding issues when assessing covariance balance in the first approach, as mentioned previously, we performed this approach as a demonstration of assessing the internal validity of an empirical study.

4.5.2.1 Regression (Sample Size = 2122)

We started with the regression model. In this approach, we did not rule out the selection bias before conducting the analysis, we controlled for all the previously mentioned covariates in the regression model. Table 4.10 presented the regression results. We observed that students' baseline measure of intent to pursue science-related research careers, biomedical majors, science identity, degree expectations and research self-efficacy were significant. The BUILD scholar program effect size was 0.202, which was smaller than those we estimated in Approach One. The interpretation of the coefficients were similar to those in the previous section, so here we left it to the readers to interpret the meaning of these significant estimates. We noticed that students who self-identified as

Table 4.10: Multiple Regression Model (N = 2122)

	Estimate	Std. Error	p value	sig.level
(Intercept)	0.068	0.366	0.853	
Scholar	0.271	0.097	0.005	**
Site: A	0.081	0.087	0.351	
Site: B	-0.129	0.073	0.079	
Site: C	-0.144	0.074	0.051	
Cohort: 2017	0.016	0.078	0.84	
Cohort: 2018	-0.059	0.067	0.374	
Cohort: 2019	0.019	0.071	0.785	
Sex: Female	0.064	0.049	0.197	
Sex: Others	0.066	0.154	0.67	
Pell	0.05	0.052	0.336	
First-gen	-0.01	0.055	0.851	
High School GPA	-0.001	0.019	0.939	
Baseline	0.249	0.016	0.000	***
Major: Biomed Natural	1.077	0.062	0.000	***
Major: Biomed Social	0.597	0.101	0.000	***
Race: Asian	0.023	0.068	0.733	
Race: Black	0.09	0.08	0.262	
Race: Hispanic	0.132	0.079	0.094	
Race: Other	-0.107	0.178	0.549	
Race: Two or more	0.144	0.071	0.044	*
Degree: Academic Terminal	0.261	0.071	0.000	***
Degree: Graduate	-0.01	0.065	0.883	
Math Training	-0.001	0.047	0.983	
Science Identity	0.273	0.045	0.000	***
Research Self-efficacy	0.073	0.025	0.003	**

Adjusted R-squared: 0.402

BUILD scholar program effect size (Hedges' g): 0.202 (SE = 0.089)

sig.level: $p < 0.001^{***}, p < 0.01^{**}, p < 0.05^{*}$

"two or more races" showed significantly stronger interests in pursuing science-related careers, compared to their "White" peers. The treatment and covariates in this model explained 40.2% of the outcome variable.

This model presented a significant effect of the BUILD scholar program on the outcome variable, after controlling for all covariates. Before drawing conclusions about the effectiveness of the program, we ran a sensitivity analysis and used the baseline intent to pursue science-related research careers and science identity as benchmarks based on the regression results. This could help us assess the potential threat of unobserved confounders that were really important to the model and might have changed the results. The explanation of the results of sensitivity analysis on unobserved confounders in empirical studies was relatively standard. To ensure the accuracy of interpretation, we used terms and languages defined in Cinelli and Hazlett (2020), and used Z to denote the confounder.

4.5.2.2 Sensitivity Analysis

As suggested by Cinelli and Hazlett (2020), we presented the minimal sensitivity analysis reporting in Table 4.11. Table 4.11 reported the estimate of the treatment effect (.271) as well as its standard error (.097) and the corresponding t-value (2.808), in the regression model with a degree of freedom of 2096 ($df = 2096$). We also reported the partial R^2 of the BUILD scholar program on the outcome variable (.4%), the robustness value ($RV_{q=1}$) for bringing the point estimate of the treatment to 0 (5.9%), and the robustness value for altering the statistical significance ($\alpha = 0.05$) of the point estimate of the treatment ($RV_{q=1,\alpha=0.05}$) to insignificant (1.8%). The partial R^2 of the BUILD scholar program on the outcome variable indicated that in an extreme scenario, even if confounders explained all remaining variation of the outcome variable, they would need to explain at least .4% of the residual variation of the treatment to bring down the estimated BUILD scholar program effect to 0. The $RV_{q=1}$ and $RV_{q=1,\alpha=0.05}$ indicated that if potential founders explained 5.9% (or 1.8% if we account for sampling uncertainty) both of the residual variation of the outcome and of the treatment, this would be sufficient to explain away the treatment effect. These quantities reflected the difficulty level of turning around the treatment effect by unobserved confounders. Although these values seemed rather small and the treatment effect did not explain much of the outcome, we needed more information to help us evaluate the likelihood of having confounders that were impactful enough to overrule the significant impact of the BUILD scholar program on the outcome variable.

Table 4.11: Sensitivity Analysis of BUILD Scholar Program on Science Career Intent

Treatment	Est.	S.E.	t-value	$R^2_{Y\sim D\|X}$	$RV_{q=1}$	$RV_{q=1,\alpha=0.05}$
BUILD scholar program	0.271	0.097	2.808	0.4%	5.9%	1.8%

Bound (1×Baseline Career Intent) : $R^2_{Y\sim Z|X,D}$ *= 11%,* $R^2_{D\sim Z|X}$ *= 0%*

Bound (1×Science Identity) : $R^2_{Y\sim Z|X,D}$ *= 1.8%,* $R^2_{D\sim Z|X}$ *= 1.1%*

To better understand the above values, we used the baseline intent to pursue science-related research careers and science identity as benchmarks for comparisons, because these variables were the most significant covariates in the model and could, to a certain extent, represent the most extreme cases. We believed that even if there were any latent confounders that we failed to detect and thus excluded in the regression analysis, due to the fact that previous literature did not at all mention covariates other than those that were already included in our model, the unobserved confounders were unlikely to be as influential as the observed strongest covariates (Cinelli & Hazlett, 2020), such as baseline science career intent and science identity, in the model. This provided the rationale for us to choose these two variables as benchmarks to provide bounds on confounding as strong as themselves. At the bottom of Table 4.11, we reported the partial R^2 of covariates with the outcome ($R^2_{Y\sim Z|X,D}$), and the partial R^2 of covariates with the treatment ($R^2_{D\sim Z|X}$).

For the variable baseline science career intent, although it explained 11% of the outcome variable, which is higher than both $RV_{q=1}$ and $RV_{q=1,\alpha=0.05}$, it did not explain anything about the treatment, indicating that this variable was not an observed confounder and it would not influence the treatment effect on the outcome variable. This finding implied that although sensitivity analysis was a post-hoc analysis on potential latent confounders, it could potentially help us confirm the existence of observed confounders as well. Even though we confirmed that this baseline variable was not a confounder, we kept it in the analysis for reference purposes.

The variable science identity was an observed confounder that explained 1.8% of the outcome variable and 1.1% of the treatment. Both of the values were not larger than $RV_{q=1}$ and $RV_{q=1,\alpha=0.05}$, indicating that confounders as strong as science identity were not sufficient to explain away the treatment effect, nor were they sufficient to change the significant conclusion of the treatment effect on the outcome variable. However, since the bound on $R^2_{D\sim Z|X}$ (1.1%) was larger than the $R^2_{Y\sim D|X}$ (.4%), an extreme confounder that explained all residual variation of the outcome variable and was as strongly associated with the BUILD scholar program assignment as the variable science identity, would be powerful enough to overrule the conclusion of the program effect on the outcome variable.

Using the statistical tool developed by Cinelli, Ferwerda, and Hazlett (2020), we further explored the potential influence of unobserved confounders if they were stronger than science identity. In Table 4.12, we presented $R^2_{D\sim Z|X}$ and $R^2_{Y\sim Z|X,D}$ of confounders that were one, two and three times as strong as baseline science career intent and science identity. The growth of partial R^2 by the strength of confounding was linear. Based on the explained proportions of the outcome and the treatment, we reported the adjusted estimations of the treatment effect on the outcome and their corresponding standard errors and t-values. We observed that the increasing strength of the baseline career intent confounding did not influence the estimates and conclusions too much. This was expected, as the baseline career intent barely correlated with the treatment variable. Increasing strength on the baseline career intent confounding, however, largely influenced the conclusions. The treatment effect estimates were brought closer and closer to 0, and a confounder that was twice as strong as the science identity was strong enough to change the estimate to be insignificant.

Table 4.12: Sensitivity Analysis (Multiple Bounds)

Bound	$R^2_{D\sim Z\|X}$	$R^2_{Y\sim Z\|X,D}$	Adj. Est.	Adj. S. E.	Adj. t-value
1 × *Baseline Career Intent*	0.000	0.110	0.246	0.091	2.703
2 × *Baseline Career Intent*	0.001	0.219	0.222	0.085	2.596
3 × *Baseline Career Intent*	0.001	0.329	0.197	0.079	2.486
1 × *Science Identity*	0.011	0.018	0.207	0.096	2.156
2 × *Science Identity*	0.023	0.036	0.143	0.096	1.492
3 × *Science Identity*	0.034	0.054	0.078	0.096	0.816

We used contour plots (Figure 4.5 and Figure 4.6) to illustrate the dynamic changes of treatment effects. In Figure 4.5, we observed that a confounder that was one, two or three times as strong as one of the two variables was not strong enough to bring the positive estimate down to 0. Figure 4.6 showed that a confounder that was two or three times as strong as science identity would bring the t-value to 1.492 and .816, which implied that the treatment effect would no longer be significant.

Cinelli et al. (2020) also provided suggestions of using sensitivity plots to analyze extreme scenarios (pp. 13-14). In our analysis, the situations were relatively extreme, since one variable (baseline career intent) barely explained any effects of the treatment, and the other one (science identity) explained a large proportion. Accordingly, we decided to report the extreme scenarios under the condition that a confounder was as strongly associated with BUILD scholar program assignment as the science identity, i.e., this confounder explained 1.1% of the treatment variable. In this scenario, if the confounder explained 31.6% (can be easily computed from $\frac{5.9\%\times 5.9\%}{1.1\%}$), it could bring the treatment effect estimate to 0; if the confounder explained 2.95% ($\frac{1.8\%\times 1.8\%}{1.1\%}$), it could bring the treatment effect estimate to be insignificant. If a confounder had a stronger association with the treatment, say, three times as strong as the science identity (3.4%),

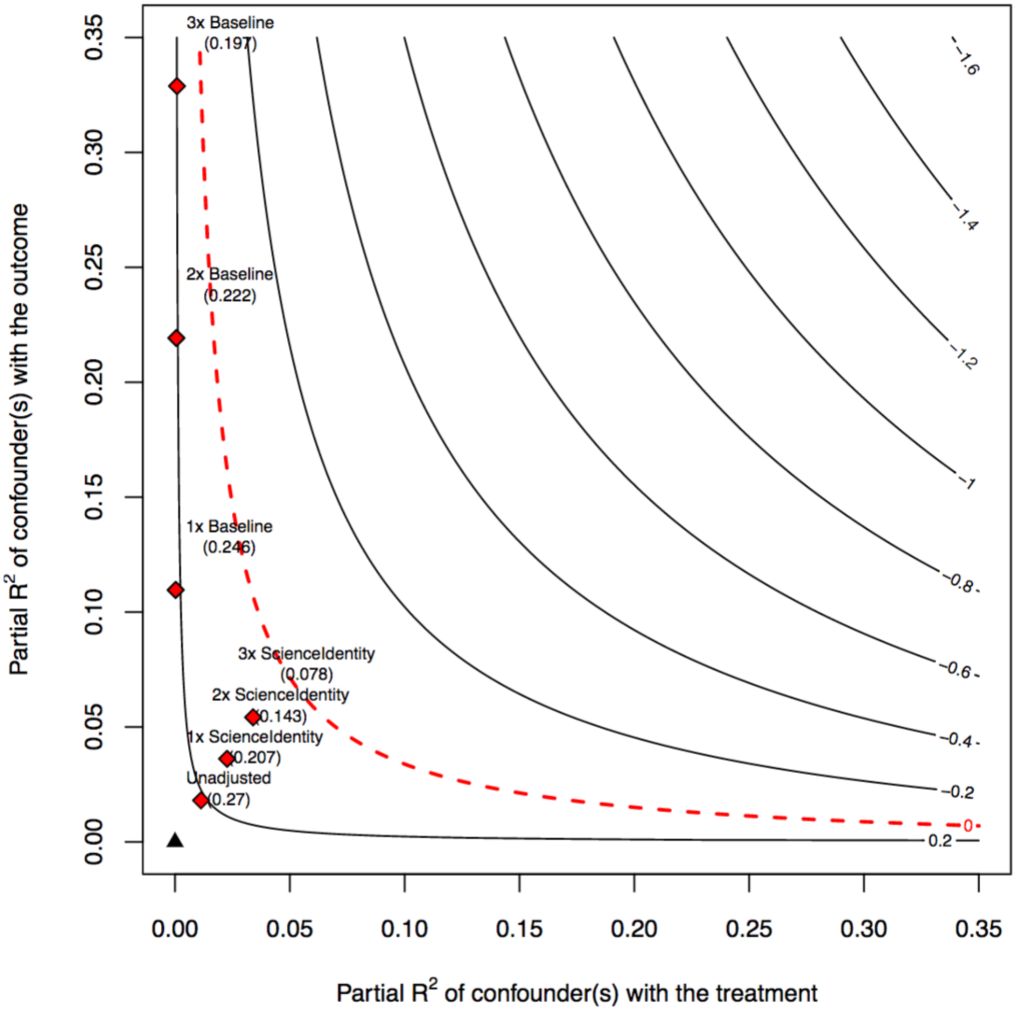


Figure 4.5: Sensitivity Analysis Contour Plots

the above two values would be 10.24% and .95%, respectively.

From previous studies, we believed that it would be unlikely to have an unobserved confounder that was as impactful as the baseline career intent to the outcome variable, since students basically responded to the same question at different time points. We also believed that unobserved confounders were unlikely to explain more of the treatment than science identity (could be explored through analysis of variances). Under these assumptions, we concluded that it was possible to have an influential unobserved confounder that could shrink the treatment effect estimate, but the confounder was quite

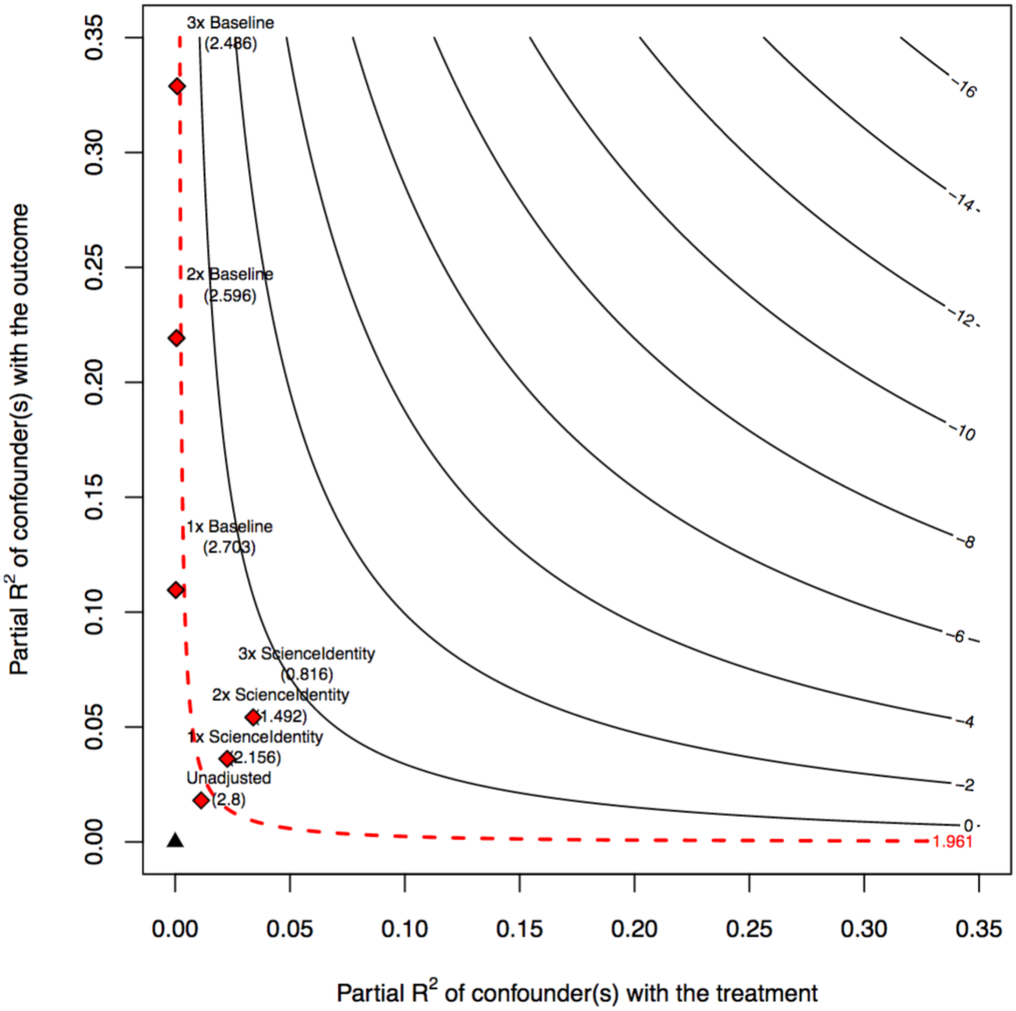


Figure 4.6: Sensitivity Analysis Contour Plots (t-value)

unlikely to be powerful enough to bring the effect to be 0, or insignificant. We were also aware of the possibility of having multiple unobserved confounders, and the previous results would be conservative for this situation (Cinelli et al., 2020). Even with that in mind, the program effect was still likely to stay positive, while the significant level might be jeopardized.

4.6 Discussion

4.6.1 Summary of Findings

In this study, we examined the effectiveness of an undergraduate diversity training program, the BUILD scholar program, on students' intent to pursue science-related research careers during students' initial stage in college. To address the research question, "Does participation in the BUILD scholar program during freshman year impact students' intent to pursue science-related research careers?" we identified 4 primary BUILD sites that provided the scholar program interventions for their first year students, and based on the program characteristics, we constructed an analytical framework (Figure 4.3), a structured and simplified application of the IEO framework (Astin, 1970a, 1970b), to evaluate potential causal relations.

Due to selection processes, BUILD scholars had higher high school GPA, higher science identity and higher research self-efficacy than the control group students prior to college. Accordingly, we utilized two approaches, one featured by multi-stage matching and the other by sensitivity analysis, to analyze BUILD program participation data and longitudinal survey data. Approach one used multi-stage matching to identify comparable control group students with the BUILD scholars and to achieve conditional ignorability through diminishing the effects of path A (Figure 4.3) prior to data analysis. Approach two implemented a post-hoc sensitivity analysis to assess the severity of the influence of observed confounders on the magnitude of effects of path B (Figure 4.3) and the likelihood of having unobserved confounders that could threaten the conclusion. Results from both approaches suggested that the BUILD scholar program positively influenced students' intent to pursue science-related research careers during students' initial stage in college. Although the BUILD scholar program effect size were less than .3 which indicated small effects, compared with other interventions in higher education, for example, interventions motioned in Sneyers and De Witte (2018), the effect sizes from the two approaches were relatively large.

In addition to answering the research question, through the first approach, we found that the baseline measure of students' science career intent, science identity, major, degree expectation, and high school GPA might influence students' intent to pursue science-related research careers. Through the second approach, we found a similar set of significant covariates – baseline measure of students' science career intent, science identity, major, degree expectation and research self-efficacy. In the second approach, students in the "two or more" race category showed stronger science career intent than their White peers. Both approaches yield similar results in that variables related to students' self-efficacy, such as baseline science career intent, science identity and degree expectation, and variables related to students' college experience, such as major and the BUILD scholar program intervention all positively influence the outcome. We did not observe significant impacts from covariates such as sites, cohort, gender identity, Pell Grant status, first generation college student status and previous math training.

4.6.2 Remarks on Methodology

We used two different approaches that emphasized different perspectives to handle confounding issues. In the first approach, we applied a multi-stage matching strategy that optimized the local similarity at the largest extent. In the second approach, we applied the regression analysis on the whole dataset and then performed sensitivity analysis to examine the plausibility of the program effects and the influence of potential confounders. Upon completing this study, we would like to comment on the analytical procedures and takeaways on methodological choices.

4.6.2.1 Notes on Matching

In this study, we designed a multi-stage matching approach that was tailored for our data and for the selection criteria. We encourage researchers to put efforts on exploring different matching procedures to identify the most suitable approach based on their own research. After confirming the potential necessity of matching (e.g., imbalanced covariates or evidence on confounders), we need to consider the following general issues for determining the matching procedures.

First of all, as shown in this study, sample size determines several aspects of matching. If the original control group sample size is large compared to that of your treatment group, it is possible to match exactly on all covariates. If the sample size is small, weighting might be a better strategy, because we do not want to lose more cases. Secondly, We are aware that ”large” and ”small” sample sizes are relative ideas, not only depending on the ratio of the original sample sizes of the treatment and control groups, but depending on the number of covariates as well. Furthermore, when involving too many covariates into the matching procedure, the curse of dimensionality makes it tricky for us to judge the balancing from checking lower dimensional summaries (Ho et al., 2007). Thirdly, the tolerance of calipers, common support and power of generalizability all set limits to the types of matching methods we are able to choose. Finally, no matching procedure should be set up as a general approach, and decisions have to be made to fit the features of the data and to maximize the balance. Ho et al. (2007) suggested that the best way to identify the optimal approach was to ”run as many as possible and choose by maximizing balance” (p. 232).

4.6.2.2 Notes on Sensitivity Analysis

Compared to the matching approach, sensitivity analysis is more standard as a statistical approach, but it requires more context knowledge for assessing the likelihood of existing unobserved confounders. The sensitivity analysis with benchmarks helps us evaluate the impact of a potential confounder, but it might be less intuitive if it is possible to have multiple confounders. The findings from this approach were drawn from the total survey participants with pre- and post-measure data, and thus the findings could be generalized to a relatively broader first year college student population, unlike in the matching approach, the results represented the treatment effect of the treated group.

4.6.2.3 Methodological Choices

With a purpose of demonstrating useful analytical approaches for researchers who might be interested in exploring causal relations in educational study, some parts of the analytical design were ”overkilled” and more than enough for answering the research question. Multiple approaches could certainly help cross-validate the findings, but in practice, for the purpose of addressing a research question in an empirical study, sometimes the analysis could be as simple as a simple regression, if the condition allowed (e.g., balanced covariates). One of the two approaches would be sufficient for most situations. As mentioned in the analysis, some procedures, such as the some stages in matching and the outcome variable adjustment could be skipped or simplified. For high-stake studies, where researchers need to provide more than enough evidence, some adjustments can be considered: setting smaller calipers (e.g., .2) or SMD thresholds (e.g., .05, as recommended by the What Works Clearinghouse, 2021) in the matching procedure, combining the two approaches to implement matching, regression analysis and sensitivity analysis without benchmarks (due to the ruling out of the partial R^2 on the treatment variable through matching), and applying rigorous statistical models such as multilevel modeling.

Every method and model has its pros and cons. For example, in addition to common support issues, some researchers (Smith & Todd, 2005) found that treatment effect estimates could be highly sensitive to both the model specification and the analytical samples. For another example, sensitivity analysis only provides information using existing covariates in the model, and assessing the likelihood of having unobserved confounders really relies on researchers’ knowledge base about the context. Sometimes, even using similar approaches, we might end up with different conclusions. Matching and sensitivity analysis only help us address the confounder issues to certain extents; we are, after all, not analyzing experimental data. We should keep in mind that ”all models are wrong but some are useful” (p. 202, Box, 1979) – we could only control as much as we could, and honestly report the procedures and results. We just have to let our readers determine the trustworthiness of our conclusions.

4.6.3 Limitations and Future Research

The foundation of data analysis in this study was that the measurement issues were addressed and measurement errors were minimized. One drawback of our study was that our outcome measure was based on a single question, which might not be reliable. We could improve the study by using scales that contained multiple items to optimize the internal reliability of the outcome measure. This stimulates us to identify potential scales that measure college students' intent to pursue science-related research careers.

In this study, although we used multiple approaches and concluded that the BUILD scholar program positively influenced students' intent to pursue science-related research careers during students' initial stage in college, one thing that we could not ignore was that the BUILD scholar program effect and the effect size were less than .3. This indicated that the program might not have enough power to motivate students to move from, say "uncertain" to "maybe yes" to pursue science-related careers, as we often discussed – statistically significant but not empirically significant (although this is really not a type I error situation). This might be due to the fact that the pre- and post-measure were only several months apart, and students barely finished one semester of their college. This implied that, in the future, we should conduct longitudinal analysis to monitor the continuity of growth.

Although results indicated potential influence of the covariates (or inputs) on the outcome variable, we needed to be careful about our interpretation. As mentioned earlier, the only thing we cared about in Figure 4.3 was the magnitude of effects of path B and the plausibility of this quantity, meaning that all our efforts in the two approaches were to ensure that we could answer the research question with certain confidence. We did not take any approach to assess if the covariates were independent with each other, and as a result, we did not know if there was any latent variable that confounded the covariates and the outcome variable. For example, the high associations between students' self-efficacy related variables and the outcome variable might be due to a latent trait that reflected a person's sense of confidence. We might get some sense through this study that certain covariates were more important. Instead of making uncertain claims, we could conduct another study that is purposefully designed to explore those covariates.

Another issue was that our ultimate goal was to identify the best practice for supporting the URGs. Although we might be able to say that the BUILD program was effective, the essential question would be – which part of the BUILD scholar program brought the significant difference. To explore the details of program activities, other forms of data collection and research practice such as site visit and case study might supplement the current research findings.

CHAPTER 5

Summary

In this dissertation, I utilized the DPC survey data to conduct three studies, demonstrated how to validate scales for a particular population, and how to measure program effectiveness. These studies could contribute to the current literature in applied measurement, and to the DPC program evaluation. Moreover, each of the studies would provide additional contributions to the methods implemented during the process.

Study one performs scale validation for a particular population, and item selection for creating short forms to make the scale a better-fit in the survey. Unlike in many other studies where researchers usually analyze the responses to one scale, in this study, we take a holistic approach. We treat the scale as a part of the survey, and adjust the scale to fit the survey instrument. As a product of this study, MCA-short-C, a 9-item short form of the original MCA was developed for measuring undergraduate faculty-student mentorship in college settings. The MCA-short-C is offered as an alternative to the long form in measuring faculty mentoring competency, particularly when 1) researchers are concerned about space constraints and/or respondent burden, 2) the participants are college faculty and students, and 3) researchers intend to explore the ″cross-rating″ of faculty mentoring from mentors' self-rating as well as mentees' rating on their mentors' competency.

Study two focuses on the response pattern of the population, and explores strategies to treat response categories prior to identifying the measurement model. This study uses an item response tree (IRTree) modeling approach to explore the possible response tendency of "N/A" responses. The "N/A" option is suggested to be included as a response category, so that participants might have more opportunities to express their real opinions to the survey items. However, how to handle "N/A" responses has rarely been studied. They are often treated as missing in analytical models, although by design, participants are provided opportunities to distinguish their response from missing. This study set an example of handling "N/A" responses, which can be a reference for DPC scale validation, and for handling other similar response options. The results indicated that we should treat such responses as "not applicable" with caution, and not all "N/A" responses could be treated as missing responses at random.

In study three, we examined the effectiveness of an undergraduate diversity training program, the BUILD scholar program, on students' intent to pursue science-related research careers during students' initial stage in college. To address the research question, "Does participation in the BUILD scholar program during freshman year impact students' intent to pursue science-related research careers?" we identified 4 primary BUILD sites that provided the scholar program interventions for their first year students, and based on the program characteristics, we constructed an analytical framework, a structured and simplified application of the IEO framework, to evaluate potential causal relations. We utilized two approaches, one featured by multi-stage matching and the other by sensitivity analysis, to analyze BUILD program participation data and longitudinal survey data. Results from both approaches suggested that the BUILD scholar program positively influenced students' intent to pursue science-related research careers during students' initial stage in college.

Using surveys to collect data for evaluating program effectiveness is a common approach in large national multi-pronged program evaluation. The three studies in this dissertation utilized the DPC survey data to address important issues in measurement, survey methods, and evaluation of program impacts. In addition to provide formative suggestions to improve the current DPC programs, these three studies demonstrated approaches for improving measurement and in large-scale surveys and using survey data to assess program impacts.

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.

Ajzen, I. (2011). *The theory of planned behaviour: Reactions and reflections.* Taylor & Francis.

Amelink, C. T., & Creamer, E. G. (2010). Gender differences in elements of the undergraduate experience that influence satisfaction with the engineering major and the intent to pursue engineering as a career. *Journal of Engineering Education*, *99*(1), 81-92. doi: https://doi.org/10.1002/j.2168-9830.2010.tb01044.x

Anctil, T. M., Hutchison, B., & Smith, C. K. (2013). Class, status, poverty, and capital: A guide to social stratification in career counseling.

Astin, A. W. (1970a). The methodology of research on college impact, part one. *Sociology of education*, 223–254.

Astin, A. W. (1970b). The methodology of research on college impact, part two. *Sociology of education*, 437–450.

Astin, A. W., & Antonio, A. L. (2012). *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education* (2nd ed.). Rowman & Littlefield Publishers.

Bandura, A. (1991). Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes*, *50*(2), 248-287. (Theories of Cognitive Self-Regulation) doi: https://doi.org/10.1016/0749-5978(91)90022-L

Berinsky, A. J., & Margolis, M. (2011). Missing voices: polling and health care. *Journal of Health Politics, Policy and Law*, *36*(6), 975–987.

Berk, R. A., Berg, J., Mortimer, R., Walton-Moss, B., & Yeo, T. P. (2005). Measuring the effectiveness of faculty mentoring relationships. *Academic medicine*, *80*(1), 66–71.

Betz, N. E., Klein, K. L., & Taylor, K. M. (1996). Evaluation of a short form of the career decision-making self-efficacy scale. *Journal of career assessment*, *4*(1), 47–57.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29–51.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual review of psychology*, *53*(1), 605–634.

Bottia, M., Stearns, E., Mickelson, R., Moller, S., & Parker, A. (2015). The relationships among high school stem learning experiences and students' intent to declare and declaration of a stem major in college. *Teachers College Record*, *117*(3), 1–46.

Box, G. E. (1979). Robustness in the strategy of scientific model building. In R. L. Launer & G. N. Wilkinson (Eds.), *Robustness in statistics* (pp. 201–236). Academic Press.

Bradbury, L. U., & Koballa Jr, T. R. (2008). Borders to cross: Identifying sources of tension in mentor–intern relationships. *Teaching and teacher education*, *24*(8), 2132–2145.

Bridges to the Baccalaureate (B2B). (n.d.). Bridges to the Baccalaureate Research Training Program (T34). *National Institute of General Medical Sciences (NIGMS)*. Retrieved from `https://www.nigms.nih.gov/research/mechanisms/pages/bridges baccalaureate.aspx`

Brown, C., Glastetter-Fender, C., & Shelton, M. (2000). Psychosocial identity and career control in college student-athletes. *Journal of Vocational Behavior*, *56*(1), 53–62.

Cai, L. (2012). Three cheers for the asymptotically distribution free theory of estimation and inference: Some recent applications in linear and nonlinear latent variable modeling. In *Current topics in the theory and application of latent variable models* (pp. 119–131). Routledge.

Campbell, D. T., & Erlebacher, A. (1970). How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. *Compensatory education: A national debate*, *3*, 185–210.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (Vol. 5, pp. 171–246). Chicago: Rand McNally.

Carlone, H. B., & Johnson, A. (2007). Understanding the science experiences of successful women of color: Science identity as an analytic lens. *Journal of Research in Science*

Teaching: The Official Journal of the National Association for Research in Science Teaching, *44*(8), 1187–1218.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research*, *1*(2), 245–276.

Cinelli, C., Ferwerda, J., & Hazlett, C. (2020). sensemakr: Sensitivity analysis tools for ols in r and stata. *Available at SSRN 3588978*.

Cinelli, C., & Hazlett, C. (2020). Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *82*(1), 39–67.

CIRP Constructs. (n.d.). Higher Education Research Institute (HERI). Retrieved from `https://heri.ucla.edu/cirp-constructs/`

CIRP Freshmen Survey. (2021). Higher Education Research Institute (HERI). Retrieved from `https://heri.ucla.edu/cirp-freshman-survey/`

Cobian, K. P., & Gutiérrez, Á. (2021). Advancing diversity, equity, and inclusion in the health services and policy research workforce: Lessons and implications from case studies on diversity, equity, and inclusion interventions – horizon scan.

Cochran, W. G., & Rubin, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, 417–446.

Cohn, D., & Caumont, A. (2016). 10 demographic trends that are shaping the us and the world. *Pew Research Center*.

Colvin, J. W., & Ashman, M. (2010). Roles, risks, and benefits of peer mentoring relationships in higher education. *Mentoring & Tutoring: Partnership in Learning*, *18*(2), 121–134.

Crisp, G., & Cruz, I. (2009). Mentoring college students: A critical review of the literature between 1990 and 2007. *Research in higher education*, *50*(6), 525–545.

Crisp, G., Nora, A., & Taggart, A. (2009). Student characteristics, pre-college, college, and environmental factors as predictors of majoring in and earning a stem degree: An analysis of students attending a hispanic serving institution. *American Educational Research Journal*, *46*(4), 924-942. doi: 10.3102/0002831209349460

Csikszentmihalyi, M. (1997). Flow and education. *NAMTA journal*, *22*(2), 2–35.

Davidson, P. L., Maccalla, N. M., Afifi, A. A., Guerrero, L., Nakazono, T. T., Zhong, S., & Wallace, S. P. (2017). A participatory approach to evaluating a national training and institutional change initiative: the build longitudinal evaluation. In *Bmc proceedings* (Vol. 11, pp. 157–169).

De Boeck, P., & Partchev, I. (2012). Irtrees: Tree-based item response models of the glmm family. *Journal of Statistical Software*, *48*(1), 1–28.

Deci, E. L., & Ryan, R. M. (1985). The general causality orientations scale: Self-determination in personality. *Journal of research in personality*, *19*(2), 109–134.

Department of Education. (2015). Science, technology, engineering, and math. Retrieved from `https://www.ed.gov/stem`

Dibenedetto, C., Easterly, R., & Myers, B. (2015, 03). Can scientific reasoning scores predict the likelihood of sbae students' intent to pursue a stem career, a career in agriculture, or plan to attend college? *Journal of Agricultural Education*, *56*, 103-115. doi: 10.5032/jae.2015.01103

don't knows (dks). (n.d.).

Duffy, R. D., & Sedlacek, W. E. (2007). The presence of and search for a calling: Connections to career development. *Journal of Vocational Behavior*, *70*(3), 590–601.

Ellwood, R., & Abrams, E. (2018). Student's social interaction in inquiry-based science education: how experiences of flow can increase motivation and achievement. *Cultural Studies of Science Education*, *13*(2), 395–427.

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

Estrada, M., Woodcock, A., Hernandez, P. R., & Schultz, P. (2011). Toward a model of social influence that explains minority student integration into the scientific community. *Journal of educational psychology*, *103*(1), 206.

Fisher, T. A., & Padmawidjaja, I. (1999). Parental influences on career development perceived by african american and mexican american college students. *Journal of*

Multicultural Counseling and Development, *27*(3), 136–152.

Fleming, M., House, M. S., Shewakramani, M. V., Yu, L., Garbutt, J., McGee, R., ... Rubio, D. M. (2013). The mentoring competency assessment: validation of a new instrument to evaluate skills of research mentors. *Academic medicine: journal of the Association of American Medical Colleges*, *88*(7), 1002.

Freedman, M. (1999). *The kindness of strangers: Adult mentors, urban youth, and the new voluntarism*. Cambridge University Press.

Funding Opportunity Announcement. (2013a). RFA-RM-13-015: NIH Coordination and Evaluation Center (CEC) for Enhancing the Diversity of the NIH-Funded Workforce Program (U54). *National Institutes of Health (NIH)*. Retrieved from `https://grants.nih.gov/grants/guide/rfa-files/RFA-RM-13-015.html`

Funding Opportunity Announcement. (2013b). RFA-RM-13-016: NIH Building Infrastructure Leading to Diversity (BUILD) Initiative (U54). *National Institutes of Health (NIH)*. Retrieved from `https://grants.nih.gov/grants/guide/rfa-files/RFA-RM-13-016.html`

Funding Opportunity Announcement. (2013c). RFA-RM-13-017: NIH National Research Mentoring Network (NRMN) Initiative (U54). *National Institutes of Health (NIH)*. Retrieved from `https://grants.nih.gov/grants/guide/rfa-files/RFA-RM-13-017.html`

Gianakos, I. (1999). Patterns of career choice and career decision-making self-efficacy. *Journal of Vocational Behavior*, *54*(2), 244–258.

Gibbs Jr, K. D., & Griffin, K. A. (2013). What do i want to be with my phd? the roles of personal values and structural dynamics in shaping the career interests of recent biomedical science phd graduates. *CBE—Life Sciences Education*, *12*(4), 711–723.

Ginther, D. K., Schaffer, W. T., Schnell, J., Masimore, B., Liu, F., Haak, L. L., & Kington, R. (2011). Race, ethnicity, and nih research awards. *Science*, *333*(6045), 1015–1019.

Gómez, R. L., Ali, A., & Casillas, W. (2014). Mentorship and the professional development of culturally responsive evaluators in the american evaluation association's graduate education diversity internship (gedi) program. *New Directions for Evalua-*

tion, *2014*(143), 49–66.

Hacking, I. (1997, August). Taking bad arguments seriously. *London Review of Books*. Retrieved from `https://www.lrb.co.uk/the-paper/v19/n16/ian-hacking/taking-bad-arguments-seriously`

Hallmarks of Success. (2020). *Diversity Program Consortium Hallmarks of Success.* Retrieved from `https://www.diversityprogramconsortium.org/pages/hallmarks_all`

Hansen, M., Cai, L., Stucky, B. D., Tucker, J. S., Shadel, W. G., & Edelen, M. O. (2014). Methodology for developing and evaluating the promis® smoking item banks. *Nicotine & Tobacco Research*, *16*(Suppl_3), S175–S189.

Harren, V. A. (1979). A model of career decision making for college students. *Journal of vocational behavior*, *14*(2), 119–133.

Hartung, P. J., Lewis, D. M., May, K., & Niles, S. G. (2002). Family interaction patterns and college student career development. *Journal of Career Assessment*, *10*(1), 78–90.

Herrera, F. A., & Hurtado, S. (2011). Maintaining initial interests: Developing science, technology, engineering, and mathematics (stem) career aspirations among underrepresented racial minority students. In *Association for educational research annual meeting, new orleans, la.*

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, *15*(3), 199–236.

Holland, J. L. (1959). A theory of vocational choice. *Journal of counseling psychology*, *6*(1), 35.

Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments*. Psychological Assessment Resources.

Holman, R., Glas, C. A., Lindeboom, R., Zwinderman, A. H., & De Haan, R. J. (2004). Practical methods for dealing with'not applicable'item responses in the amc linear disability score project. *Health and quality of life outcomes*, *2*(1), 1–11.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185.

Hurtado, S., Eagan, M. K., Cabrera, N. L., Lin, M. H., Park, J., & Lopez, M. (2008). Training future scientists: Predicting first-year minority student participation in health science research. *Research in Higher Education*, *49*(2), 126–152.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jacobi, M. (1991). Mentoring and undergraduate academic success: A literature review. *Review of educational research*, *61*(4), 505–532.

Jeon, M., & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavior research methods*, *48*(3), 1070–1085.

Lavigne, G. L., & Vallerand, R. J. (2010). The dynamic processes of influence between contextual and situational motivation: A test of the hierarchical model in a science education setting. *Journal of Applied Social Psychology*, *40*(9), 2343-2359. doi: https://doi.org/10.1111/j.1559-1816.2010.00661.x

Lent, R. W. (2005). A social cognitive view of career development and counseling.

Lent, R. W., & Brown, S. D. (2006). Integrating person and situation perspectives on work satisfaction: A social-cognitive view. *Journal of vocational behavior*, *69*(2), 236–247.

Lent, R. W., Brown, S. D., & Hackett, G. (1994). Toward a unifying social cognitive theory of career and academic interest, choice, and performance. *Journal of vocational behavior*, *45*(1), 79–122.

Lent, R. W., Brown, S. D., & Hackett, G. (2000). Contextual supports and barriers to career choice: A social cognitive analysis. *Journal of counseling psychology*, *47*(1), 36.

Lin, Y.-n., & Hsu, A. Y.-p. (2012). Peer mentoring among doctoral students of science and engineering in taiwan. *Asia Pacific Education Review*, *13*(4), 563–572.

Maccalla, N. M., Gutierrez, A., Zhong, S., Wallace, S. P., & McCreath, H. E. (2020). Technical report: Evaluation of post-secondary student outcomes: Underrepresented (URG) and well-represented (WRG) group variable construction in the enhance diversity study using the november 2019 nih guidelines. Retrieved from `https://www.diversityprogramconsortium.org/files/view/docs/CEC_Techinical_Report_URG_WRG_Group_Variable_Construction_Aug2020.pdf`

MacLachlan, A. J. (2012). Minority undergraduate programs intended to increase participation in biomedical careers. *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine*, *79*(6), 769-781. doi: https://doi.org/10.1002/msj.21350

Maximizing Access to Research Careers (MARC). (n.d.). Maximizing Access to Research Careers (MARC) Awards (T34). *National Institute of General Medical Sciences (NIGMS).* Retrieved from https://www.nigms.nih.gov/training/MARC/Pages/USTARAwards.aspx

McCreath, H. E., Norris, K. C., Calderón, N. E., Purnell, D. L., Maccalla, N. M., & 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. In *Bmc proceedings* (Vol. 11, pp. 15–26).

McGee Jr, R., Saran, S., & Krulwich, T. A. (2012). Diversity in the biomedical research workforce: developing talent. *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine*, *79*(3), 397–411.

Mishkin, H., Wangrowicz, N., Dori, D., & Dori, Y. J. (2016). Career choice of undergraduate engineering students. *Procedia - Social and Behavioral Sciences*, *228*, 222-228. (2nd International Conference on Higher Education Advances,HEAd'16, 21-23 June 2016, València, Spain) doi: https://doi.org/10.1016/j.sbspro.2016.07.033

NIH: NOT-OD-20-070. (2020). *NOT-OD-20-070: Ruth L. Kirschstein National Research Service award (NRSA) Stipends, Tuition/Fees and other BUDGETARY levels effective for fiscal year 2020.* U.S. Department of Health and Human Services. Retrieved from https://grants.nih.gov/grants/guide/notice-files/NOT-OD-20-070.html

NRMN. (n.d.). National Research Mentoring Network (NRMN). *NRMN*. Retrieved from https://nrmnet.net/about-nrmn/

NRMN. (2019). National Research Mentoring Network (NRMN). *Diversity Program Consortium (DPC).* Retrieved from https://www.diversityprogramconsortium.org/pages/nrmn

Oldendick, R. W. (2008). Response alternatives. In *Encyclopedia of survey research methods*

(Vol. 1, pp. 750–751). SAGE Thousand Oaks, CA.

Oldendick, R. W. (2012). Survey research ethics. In L. Gideon (Ed.), *Handbook of survey methodology for the social sciences* (pp. 23–35). New York, NY: Springer New York. doi: 10.1007/978-1-4614-3876-2_3

Pajares, F., & Schunk, D. H. (2001). Self-beliefs and school success: Self-efficacy, self-concept, and school achievement. *Perception*, *11*(2), 239–266.

Pascarella, E. T., & Staver, J. R. (1985). The influence of on-campus work in science on science career choice during college: A causal modeling approach. *The Review of Higher Education*, *8*(3), 229-245. doi: 10.1353/rhe.1985.0019

Pearl, J., & Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.

Peterson, S. L. (1993). Career decision-making self-efficacy and institutional integration of underprepared college students. *Research in higher education*, *34*(6), 659–685.

Pfund, C., House, S., Spencer, K., Asquith, P., Carney, P., Masters, K. S., ... Fleming, M. (2013). A research mentor training curriculum for clinical and translational researchers. *Clinical and translational science*, *6*(1), 26–33.

Pfund, C., House, S. C., Asquith, P., Fleming, M. F., Buhr, K. A., Burnham, E. L., ... others (2014). Training mentors of clinical and translational research scholars: a randomized controlled trial. *Academic medicine: journal of the Association of American Medical Colleges*, *89*(5), 774.

Quinn, H. O. (2014). Bifactor models, explained common variance (ecv), and the usefulness of scores from unidimensional item response theory analyses.

Rask, K. (2010). Attrition in stem fields at a liberal arts college: The importance of grades and pre-collegiate preferences. *Economics of Education Review*, *29*(6), 892–900.

Research Training Initiative for Student Enhancement (RISE). (n.d.). Research Training Initiative for Student Enhancement (RISE) Program. *National Institute of General Medical Sciences (NIGMS)*. Retrieved from `https://www.nigms.nih.gov/training/RISE/Pages/default.aspx`

Rickles, J. H., & Seltzer, M. (2014). A two-stage propensity score matching strategy for treatment effect estimation in a multisite observational study. *Journal of Educational and Behavioral Statistics*, *39*(6), 612–636.

Riegle-Crumb, C., & Morton, K. (2017). Gendered expectations: Examining how peers shape female students' intent to pursue stem fields. *Frontiers in Psychology*, *8*, 329. doi: 10.3389/fpsyg.2017.00329

Robnett, R. (2013). The role of peer support for girls and women in stem: Implications for identity and anticipated retention. *International Journal of Gender, Science and Technology*, *5*(3), 232–253. Retrieved from `http://genderandset.open.ac.uk/index.php/genderandset/article/view/299`

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688.

Sahin, A., Ekmekci, A., & Waxman, H. C. (2017). The relationships among high school stem learning experiences, expectations, and mathematics and science efficacy and the likelihood of majoring in stem in college. *International Journal of Science Education*, *39*(11), 1549-1572. doi: 10.1080/09500693.2017.1341067

Samejima, F. (1979). *A new family of models for the multiple-choice item.* (Tech. Rep.). TENNESSEE UNIV KNOXVILLE DEPT OF PSYCHOLOGY.

Schofield, L. S. (2015). Correcting for measurement error in latent variables used as predictors. *The annals of applied statistics*, *9*(4), 2133.

Schultz, P. W., Hernandez, P. R., Woodcock, A., Estrada, M., Chance, R. C., Aguilar, M., & Serpe, R. T. (2011). Patching the pipeline: Reducing educational disparities in the sciences through minority training programs. *Educational evaluation and policy analysis*, *33*(1), 95–114.

Seibert, S. (1999). The effectiveness of facilitated mentoring: A longitudinal quasi-experiment. *Journal of vocational behavior*, *54*(3), 483–502.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference.

Smith, J. A., & Todd, P. E. (2005). Does matching overcome lalonde's critique of nonexperimental estimators? *Journal of econometrics*, *125*(1-2), 305–353.

Smits, D. J., De Boeck, P., & Vansteelandt, K. (2004). The inhibition of verbally aggressive behaviour. *European Journal of Personality*, *18*(7), 537–555.

Sneyers, E., & De Witte, K. (2018). Interventions in higher education and their effect on student success: a meta-analysis. *Educational Review*, *70*(2), 208–228.

Stone, C., Van Horn, C., & Zukin, C. (2012). Chasing the american dream: Recent college graduates and the great recession. *John J. Heldrich Center for Workforce Development, Rutgers University*.

Stuart, E. A., Lee, B. K., & Leacy, F. P. (2013). Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *Journal of clinical epidemiology*, *66*(8), S84–S90.

Super, D. E. (1953). A theory of vocational development. *American psychologist*, *8*(5), 185.

Super, D. E. (1980). A life-span, life-space approach to career development. *Journal of vocational behavior*, *16*(3), 282–298.

Super, D. E. (1990). A life-span, life-space approach. career choice and development. *Brown, D. & Brooks, L. San Fransisco: Jossey-Bass Publisher*.

Sweeney, J., & Villarejo, J. K. M. (2013). Influence of an academic intervention program on minority student career choice. *Journal of College Student Development*, *54*, 534 - 540.

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, *49*(4), 501–519.

Thomson, W. (1889). Nature series. popular lectures and addresses.(a lecture delivered at the institution of civil engineers, may 3, 1883) vol. 1. *Macmillan and Co London*.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of educational psychology*, *16*(7), 433.

Thurstone, L. L. (1947). Multiple-factor analysis; a development and expansion of the vectors of mind.

Valantine, H. A., & Collins, F. S. (2015). National institutes of health addresses the science of diversity. *Proceedings of the National Academy of Sciences*, *112*(40), 12240–12242.

Wang, X. (2013). Why students choose stem majors: Motivation, high school learning, and postsecondary context of support. *American Educational Research Journal*, *50*(5), 1081–1121.

What Works Clearinghouse. (2021). Wwc procedures handbook. *WWC — Handbooks and Other Resources*. Retrieved from `https://ies.ed.gov/ncee/wwc/Handbooks`

Willis, R. J., & Rosen, S. (1979). Education and self-selection. *Journal of political Economy*, *87*(5, Part 2), S7–S36.

Wright, S. (1921). Correlation and causation. *Jour. Agric. Res*, *20*, 557–585.

Young, D., Fraser, B., & Woolnough, B. (1997). Factors affecting student career choice in science: An australian study of rural and urban schools. *Research in Science Education*, *27*, 195-214. doi: https://doi.org/10.1007/BF02461316

Zhang, C. (2016). *Generalized irtree models of children's analogical reasoning processes* . Leiden University.

Zhong, S. (2016). *The effectiveness of a peer mentorship program: A mixed methods study* . University of Southern California.

Zhong, S., Maccalla, N. M. G., & Jeon, M. (2020). Technical Report: Short-Form of the Mentoring Competency Assessment (MCA-short). *Diversity Program Consortium.* Retrieved from `https://www.diversityprogramconsortium.org/briefs/pages/MCA_short`