HEISER, CIJI ANN, Ph.D. Examining Cultural Responsiveness and Invariance in the National Survey of Student Engagement for First-generation College Students. (2020) Directed by Dr. John Willse and Dr. Jill Chouinard. 186 pp.

In higher education, there is a gap in our collective understanding of how outcomes data collected with standardized instruments and used to respond to accountability demands, is accurate and trustworthy for first-generation college students. Research in culturally responsive evaluation and measurement on equivalence across groups highlights that quantitative measures, standardized on dominant populations, lack cultural responsiveness and equivalence. Failing to critically examine if a measure is culturally responsive and invariant upholds normative assumptions that all student experiences and knowledge are captured accurately. A prolific measure of outcomes in higher education, the National Survey of Student Engagement, has gone unexamined for cultural responsiveness and invariance for first-generation college students. The purpose of this research was to identify and employ strategies to determine to what extent the National Survey of Student Engagement is culturally responsive and invariant for first-generation college students. I used a parallel convergent study design to investigate this problem. First, I conducted a critical examination of the empirical literature in culturally responsive evaluation and measurement to identify core considerations for determining if a measure is culturally responsive and invariant. Second, I conducted a multi-group confirmatory factor analysis to establish configural, metric, and scalar invariance. From study one, two core considerations emerged including attention to voice and establishing cultural relevance and invariance. Study two showed that the National Survey of Student Engagement is invariant at the configural, metric, and scalar levels. Taking the results of

the two studies together, one can determine that the National Survey of Student Engagement is culturally responsive and invariant for first-generation college students when compared to continuing-generation college students in this study.

EXAMINING CULTURAL RESPONSIVENESS AND INVARIANCE IN THE NATIONAL SURVEY OF STUDENT ENGAGEMENT FOR FIRST-GENERATION COLLEGE STUDENTS

by

Ciji Ann Heiser

A Dissertation Submitted to
the Faculty of The School of Education at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2020

Approved by

_____
Committee Co-Chair

_____
Committee Co-Chair

To Sid, Joann, Evan, Clark, and Sidney

APPROVAL PAGE

This dissertation written by CIJI ANN HEISER has been approved by the following committee of the faculty of The School of Education at The University of North Carolina at Greensboro

Committee Co-Chair _____

Committee Co-Chair _____

Committee Members _____

_____

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

ACKNOWLEDGMENTS

Throughout the course of this doctoral journey, my life has changed in a number of ways. During my first semester, I found out Evan and I were pregnant with our first baby, Clark Oliver Heiser. After Clark, I was promoted at work, eventually left UNC-CH, and moved to Michigan for a new job and a new home. In my last year of coursework I lost the two most foundational people in my life, Joann Lee and Sidney Lee, my grandparents. They were here when I started this journey; they were proud of me then and would be so proud now. After these losses, I experienced a miscarriage. The compounding loss was profound. Following my proposal, while writing my dissertation, I had Sidney Evie Ann Heiser. Shortly after she was born, COVID-19 hit, and I spent the first several weeks right after acquiring my data, wondering if I would have a job, if my family would be insured, if my husband would keep his job, if my kids and family would be safe. This journey has been long, hard, and would have been impossible with a different set of faculty, friends, and colleagues.

First, I want to thank the faculty in this program. When I drove to campus to meet with Dr. Penfield before applying to the program, he told me it was possible to complete the M.S. /Ph.D. and work fulltime. Six years later, he was right. Each of you have supported me as a student, professional, and person over the past six years. John, your candid conversations, open feedback, and kindness got me to the end of this journey. Jill, Bob, and Ayesha, your flexibility in letting me join class remotely, long before that was the only option, kept me in the program and made me an alumna of UNCG. Jill and Bob,

your courses on culturally responsive evaluation and structural equation modeling lit a fire I did not know I had. Thank you. Ayesha, your gentle compassion and kindness reminded me that it was okay to move at a reasonable pace, take space, and breathe. Those reminders are rare in my world; I am grateful for your calm compassion.

Second, thank you to the most supportive friends and colleagues a woman could ask for. To Amy, Allan, and Diane, thank you for supporting me on this journey. Thank you, Krista, Sarah, and Kate for support in navigating this journey. Dr. Jody Brylinsky, for getting my regalia wardrobe started – thank you. Nicole, thank you for letting me join you on your journey and teaching me along the way.

To my Aunt Barb, thank you for being the best cheerleader I could have ever asked for. You always ask about my research, give me ideas, encourage my brainstorming, and remind me of what I am capable. I love you. Thank you. To Nancy and Chuck Heiser for spending countless hours with Evan, Clark, and Sidney while I worked on school, for making us dinner, helping with jobs around our house, supporting our move from North Carolina to Michigan - thank you. I cannot wait to celebrate with you. To Susan, for coming for long weekends and making weeks' worth of meals that got us through the hardest days of balancing two doctorates, two jobs, and a baby. To Alex, for snuggling with baby Clark and giving us space to rest. To David, for the gentle reminders that we can do some things, but not all things, and not all at once. To John and Amy Heiser, thank you for never asking questions about when I would be "done," for being such good role models of being healthy, and for encouraging fun and relaxation. To Lucy, thank you for asking questions about item response theory, looking at my books

with me, and helping me highlight my notes. Sloane, thanks for being the best cousin Clark-O could ask for and spending so much time with him while I write.

To my son, thank you for sitting next to me while I worked through SAS and R, and smiling at me when I struggled. You are the best study buddy, and dragon rider, a mom could ask for. Thank you for your endless patience with me throughout your five years and encouraging me with your question of "When will you be Dr. Mom?" Sidney, you are almost a year old: thank you for sleeping through the nights and handling teething like a little champion. I'm so grateful for your giant smiles, constant giggles, and the joy you bring to our lives.

Finally, to my husband. Evan, who completed his Ed.D. while I was in this program, there really are no words. I'm so grateful for every meal you cooked, load of laundry you did, vet appointment you made for the dog, the grocery shopping, and every time you made plans for you and the kids so I could get this work done. I'm grateful for you and our partnership. I love you, thank you. We did it.

TABLE OF CONTENTS

Page

LIST OF TABLES ix

LIST OF FIGURES x

CHAPTER

I. INTRODUCTION 1

Background 2
Culturally Responsive Evaluation 11
Measurement across Cultural Groups 12
Validity 14
Problem Statement 15
Study Purpose and Research Questions 16
Significance 18
Manuscript Organization 19

II. LITERATURE REVIEW 20

Culturally Responsive Evaluation 21
Measurement 35
Validity 41
A Case Example 47
Culturally Responsive Evaluation as a Theoretical Lens for this Study 58
Summary of the Literature 62

III. METHODOLOGY 65

Research Design 65
Study One: Critical Literature Review 70
Study Two: Multi-Group Confirmatory Factor Analysis 73
Data Quality 81
Researcher Positionality 83

IV. RESULTS: STUDY ONE 85

Review of the Literature 85
Review Strategy and Analysis 87
Descriptive Analysis 88
Discussion 101

Critical Considerations....106

V. RESULTS: STUDY TWO....110

Sample Demographics....110
Multi-group Confirmatory Factor Analysis....114
Discussion....122

VI. DISCUSSION AND CONCLUSIONS....128

Connections across Study One and Two....128
Contributions to Research on Evaluation and Higher Education....131
Limitations....132
Future Research: What is Next?....134
Conclusions....136

REFERENCES....138

APPENDIX A. CRITICAL EXAMINATION OF THE LITERATURE....160

APPENDIX B. ANALYSIS RESULTS....179

LIST OF TABLES

Page

Table 1: Demographics for First-generation and Continuing-generation College Students (Redford & Hoyer, 2017) 6

Table 2: Threats and Justifications for Multicultural Validity (Kirkhart, 2013) 44

Table 3: Academic Progress (Cataldi et al., 2018) 54

Table 4: Summary of Research Questions, Methods, and Data Sources 66

Table 5: NSSE Themes, Factors, and Items 74

Table 6: Reliability of NSSE Indicators (NSSE, 2019) 77

Table 7: Reliability, Means, and Variance by Factor and Group 112

Table 8: Fit Indices for the Baseline Model 115

Table 9: Fit Indices for Configural, Metric, and Scalar Models 117

Table 10: Summary of Standardized Loadings Below .7 In At Least One Group 120

Table 11: Effect Size for Each Factor 121

Table 12: Matrix from Critical Examination of the Literature 160

Table 13: Configural Analysis Results 179

Table 14: Metric Analysis Results 181

Table 15: Scalar Analysis Results 184

LIST OF FIGURES

Page

Figure 1: Cultural Responsiveness and Invariance as Evidence of Multicultural Validity 46

Figure 2: Culturally Responsive Evaluation as a Theoretical Lens 62

Figure 3: Research Design 68

Figure 4: Configural Invariance 79

Figure 5: Metric Invariance 80

CHAPTER I

INTRODUCTION

Designed during the Colonial era, the founders of American higher education institutions adopted the English, Oxford, and Cambridge model of education; and the students served in that model were wealthy, White, young men (Thelin, 2019). Advances in educational access initiatives have created a significant shift in the demographic population of higher education. Institutions are evolving, but not at the same rate as the student population. As a result, the systems and structures historically underrepresented students interact with still reflect a colonial heritage. Significant differences in retention and graduation rates across marginalized groups (National Center for Education Statistics, 2016) suggest such groups have a different experience in, and are served differently by, higher education institutions. Prior to graduation, students engage in the post-secondary environment in ways that foster academic achievement and serve as markers of progress towards completion. Standardized surveys are the dominant method for measuring engagement outcomes (Kuh, Jankowski, Ikenberry, & Kinzie, 2014). Practitioners in higher education do not know if the standardized surveys used to measure outcomes reinforce dominant White norms or authentically capture outcome achievement for historically underrepresented students. Research in culturally responsive evaluation and measurement suggests standardized measures may not be culturally responsive or invariant across diverse populations (Chouinard & Cousins, 2009; Frierson, Hood,

Hughes, & Thomas, 2010; Hambleton, Merenda, & Spielberger, 2005). Failing to critically examine standardized measures for cultural responsiveness and invariance has consequences for equity and justice. This study aims to identify how to ensure standardized measures are culturally responsive and invariant using the National Survey of Student Engagement in higher education as a case example.

Background

The racial and socio-economic profile of students advancing from K-12 education to higher education has changed substantially over time. Higher education has become more accessible to students who have been historically underrepresented in this context; however, persistent disparities in educational outcomes suggests that the ways in which underrepresented students experience higher education is fundamentally different than their peers. First-generation college students are the first in their family to attend higher education. First-generation college students are racially diverse, socio-economically diverse, and have shown disparate outcomes in higher education (Cataldi et al., 2018; Gibbons et al., 2019; Pike & Kuh, 2005; Redford & Hoyer, 2017). Disparate outcomes across groups are notable because at the same time higher education has become more diverse, demands of accountability for student success and quality education persist. Outcome measurement has become a critical way to provide evidence of educational impact and respond to calls for accountability. The National Survey of Student Engagement (NSSE) is a pervasive measure of outcomes in higher education. This survey is designed to measure student engagement in purposeful activities linked to persistence and completion.

Calls for assessment approaches that are culturally responsive, leverage data to address inequities, and improve outcomes for historically underrepresented populations have emerged in response to an increasingly diverse population and the continued emphasis on outcome assessment (Montenegro & Jankowski, 2017; Zerquera, Reyes, Pender, & Abbady, 2018). Given that such calls are new in higher education, little is known about how to effectively implement such practices. However, lessons can be learned from two areas outside of higher education assessment: the study of culturally responsive evaluation and measurement across cultural groups. Culturally responsive evaluation (CRE) is a transformative approach rooted in social justice principles that centers culture and cultural ways of knowing in the evaluation process to unearth inequities and restore justice (Hood et al., 2005; 2015). CRE locates culture, and the lived experiences of participants, at the core of the evaluation process (Frierson, Hood, & Hughes, 2002). Measurement theory describes the study of measurement using applied statistics to improve existing measures and better develop new measures (Allen & Yen, 1979). Research in measurement frequently examines instruments for fairness, bias, and equivalence as a prerequisite to making valid group comparisons. Both of these professions and bodies of literature have discussed challenges associated with the use of quantitative measures across culturally diverse populations.

The National Survey of Student Engagement (NSSE) is a predominant measure of outcomes that indicate successful progress towards graduation. First-generation college students, as historically underrepresented students, have different experiences and outcomes than their majority peers (Cataldi et al., 2018; Gibbons et al., 2019; Pike &

Kuh, 2005; Redford & Hoyer, 2017). Standardized measures designed and tested without consideration for historically underrepresented students may not provide accurate or trustworthy evidence reflective of these experiences (Frierson et al., 2010). While NSSE has been developed and tested for over twenty years, NSSE has not been examined for cultural responsiveness or invariance for first-generation college students. This study draws on the empirical literature to examine how standardized measures can be identified as culturally responsive and empirically tests the extent to which NSSE is invariant for first-generation and continuing-generation college students.

Diversity in Higher Education

Many of the most prominent American higher education institutions were founded during the Colonial era and modeled after institutions such as Oxford or Cambridge. Students of this era were among the most elite in society, namely White, affluent sons of businessmen (Thelin, 2019). During this initial era, philanthropy and religion influenced higher education environments to focus on the transformation of "Christian gentlemen" into "gentlemen scholars" (Thelin, 2019). Initial attempts at expanding student diversity began with the enrollment of Native American men in an effort to convert them to Christianity, often with disastrous consequences (Thelin, 2019). Significant changes in the student body and educational experience did not begin until the 1860s. At the time, institutions had been developed for the purpose of educating women, and separately African Americans, and the educational experience broadened to include meaningful co-curricular engagement activities such as clubs, sports teams, and debate teams (Thelin, 2019). The history of American higher education is not a welcoming nor inclusive one

and, in some cases, is a history of harm. Since the 1800s, education from kindergarten to post-secondary has become increasingly diverse across institutions.

A report on *The Condition of Education 2018* provides evidence of the dramatic changes in diversity in American higher education. In 2016, 19 percent of children lived in poverty, 10 percent of children lived with a parent who had not completed high school (McFarland et al., 2018). Of students enrolled in K-12 education in the U.S., 24 percent were eligible for free or reduced-priced lunches and attended high-poverty schools; percentages for Hispanic, Black, American Indian/Alaska Native, and Pacific Islander students were higher than the national average, in some cases by as much as 20 percent (McFarland et al., 2018). At the same time, more students were successfully graduating from K-12 schools, with 84 percent of students graduating high-school four years after starting the ninth grade (McFarland et al., 2018, p. 35). Several of these trends continued into the post-secondary education environment.

Students who are the first in their family to attend higher education, referred to as first-generation college students, have emerged as a population of study, and reflect the growing diversity in higher education (Cataldi et al., 2018; Redford & Hoyer, 2017; Whitley et al., 2018). In 2012, 24 percent of students were first-generation college students (Redford & Hoyer, 2017). Given the population higher education was originally designed to serve, racially and socioeconomically diverse students are historically underrepresented and underserved. As shown in Table 1, first-generation college students are often racially diverse and from a low socio-economic background (Whitley et al., 2018).

Table 1: Demographics for First-generation and Continuing-generation College Students (Redford & Hoyer, 2017)

Demographic Variable	First-generation College Students	Continuing-generation College Students
Hispanic or Latino	27%	9%
Black or African American	14%	11%
Household income between $20,001-$50,000	50%	23%
Household income between $20,000 or less	27%	6%

First-generation college students advance from K-12 to higher education with inequitable resources related to the cost of education, food security, housing security, and tacit knowledge of higher education passed down from their parents (Pascarella et al., 2004; Pike & Kuh, 2005), all of which influences their capacity to engage, persist, and graduate at similar rates as their continuing-generation peers.

Accountability and Quality in Higher Education

In the history of American higher education, the year 1900 marked the onset of the development of criteria from which to determine the quality of institutions, differentiating between "great American universities" and "standard American universities" (Thelin, 2019, p. 111). Fourteen university presidents met and formed the Association of American Universities to respond to concerns about education standards, which they did by forming the College Entrance Examination Board, credited with establishing criteria for "ratings, rankings, and reputations" (Thelin, 2019, p. 147). More recently, discussions of quality in higher education are driven by national rankings, calls for accountability, and student outcomes. The most recognizable ranking system in

American higher education is that of the *U.S. News and World Report*, which uses a number of criteria to create ranked lists of universities (e.g., the top public institutions, best liberal arts colleges) and claims to help students make choices about which institution is the best fit for their needs. (*U.S. News and World Report*, 2019). Calls for accountability are related to calls for quality and often focus on retention, graduation, and post-graduate outcomes. Outcomes assessment has become a mechanism for responding to calls for accountability and serves as a measure of quality of student learning and engagement independent of rankings.

The release of the *Spellings Commission Report* (2006) marked a significant call for accountability in higher education and challenged institutions to demonstrate accountability, stating: "Colleges and universities must become more transparent about cost, price, and student success outcomes, and must willingly share this information with students and families" (p. 4). Bresciani, et al., (2009) reflected on demands for accountability, writing that institutions of higher education "currently face a mix of accountability demands, accreditation standards, and outcomes-based assessment of student learning. State and federal governments continue to question whether institutions of higher education actually produce the learning that has for centuries been assumed" (p. 12). Measuring outcomes identifies what students know and can do given their participation in curricular and co-curricular activities across higher education environments and allows institutions to respond to internal and external challenges around the quality of student education (Bresciani et al., 2009).

Outcomes in Higher Education

While the student populations enrolled across higher education institutions is increasingly diverse, clear differences across groups in the attainment of outcomes related to persistence, educational progress prior to completion, and graduation continues. Data collected by the National Center for Education Statistics (2016) showed graduation rates for White students entering four-year public institutions as 38.5 percent, while Black students graduated after four years at 18.1 percent. In 2006, 33 percent of first-generation college students left their institutions without returning, compared to 14 percent of their peers whose parents had a bachelor's degree (Cataldi et al., 2018). Although graduation in four years is the definitive student success outcome in higher education, a number of other indicators demonstrate student success towards completion, including academic outcomes (e.g., faculty interactions, study skills), and engagement outcomes (e.g., campus engagement, discussions with diverse peers). Evidence suggests that students from historically underrepresented backgrounds in higher education demonstrate differences in both academic outcomes (National Center for Education Statistics, 2016) and engagement outcomes (Kuh et al., 2008). In a climate of accountability, national survey measures of student outcomes have grown. One study found that national surveys are the primary way in which outcomes are measured in higher education (Kuh et al., 2014). Although the use of standardized surveys as outcome measures has grown, the quality of these measures for use with underserved populations, specifically for first-generation college students, is largely unexamined.

The National Survey of Student Engagement (NSSE) is one of the most prolific measures of student outcomes. NSSE is one of the most well researched and rigorously tested surveys in higher education (Campbell & Cabrera, 2011; Kuh, 2008, Kuh, 2009; LaNasa et al., 2009; Ouimet et al., 2004; Porter, 2011) and has been used by over 1,600 colleges and universities measuring approximately six million students since 2000 (National Survey of Student Engagement, 2019). One of the reasons for NSSE's popularity is that it emerged as an avenue to measure quality independent of "ratings, rankings, and reputations" (Thelin, 2019, p. 147) and focused on aspects of the student experience and the quality of these experiences, that institutions could concretely influence (Campbell & Cabrera, 2011). In response to compounding critiques (Campbell & Cabrera, 2011; LaNasa, et al., 2009; Porter, 2011), NSSE released an updated version of the survey in 2013 with an accompanying psychometric profile offering evidence of reliability and validity. The 2013 NSSE survey moved away from measuring five benchmarks towards the measure of ten engagement indicators designed to capture student and institutional efforts related to engagement in quality educational environments which support student completion (National Survey of Student Engagement, 2018). Multiple studies have examined the psychometric properties of NSSE and even more studies have used NSSE data to draw conclusions about marginalized populations or examine the impact of engagement on student outcomes. To date, no studies have examined if NSSE is culturally responsive or if NSSE is invariant for first-generation college students. The prevailing assumption for this instrument, normed on a predominantly White population, is that NSSE serves as an accurate and

trustworthy measure for historically underrepresented, first-generation college students. Examining this assumption creates an opportunity to establish multicultural validity and advance equity.

Cultural Responsiveness

Individuals develop in contexts shaped by culture, which shapes how learning occurs and the rules for demonstrating learning (Hughes, Seidman, & Williams, 1993). Culturally responsive approaches prioritize the thoughtful inclusion of contextual and demographic variables, address the influences of power and institutional racism, center minoritized ways of knowing and meaning making, and work to reduce marginalization and inequities (Bledsoe & Donaldson, 2015; Mertens, 2010; SenGupta et al., 2004). As different cultures value different ways of sharing their experiences, implications extending to ways in which students learn and demonstrate achievement (Maki, 2010) need to be re-considered. Surveys are used to respond to accountability claims in higher education and need to be critically examined for cultural responsiveness and invariance.

Within the climate of accountability and increasing diversity in higher education, calls for approaches that are culturally responsive, improve outcomes for historically underrepresented populations, and address educational inequities have emerged (Montenegro & Jankowski, 2017; Zerquera et al., 2018). Montenegro and Jankowski (2017) from the National Institute of Learning Outcomes Assessment (NILOA), published a white paper in 2017 on the topic of culturally responsive assessment in higher education. The authors describe cultural responsiveness as “mindful of the student populations the institution serves in ways that bring students into the assessment process

including the development and use of tools appropriate for measuring student outcomes" (Montenegro & Jankowski, 2017, p. 10). They call for alignment of assessment practices with practices that better capture the experiences of marginalized students and explain that, "by being mindful of how culture affects students' meaning-making process, cognition, and demonstrations of learning, we can better understand and appreciate the learning gains that students make" (Montenegro & Jankowski, 2017, p. 13). The current study argues that mindfulness should extend to examining the assumption that measures, standardized on Western populations, produce trustworthy and accurate data reflective of historically underrepresented and underserved student experiences. The call for cultural responsiveness is emerging in higher education, but specific attention to the measurement of student outcomes using tools that are culturally responsive and invariant is under-researched.

Culturally Responsive Evaluation

To redistribute power, expose, and address inequities, culturally responsive evaluation (CRE) focuses attention on culture throughout the evaluation process in designing the evaluation, selecting and engaging stakeholders, identifying the purpose of the evaluation, considering methods, and collecting, analyzing and reporting the data (Frierson et al., 2002). Foundational literature in CRE positions quantitative, qualitative, and mixed-method approaches to data collection and analysis as values-latent, shaped by context and culture, and as requiring decisions shaped by evaluator positionality with implications for equity and justice (American Evaluation Association, 2011; Chouinard & Cram, 2020; Frierson et al., 2002; Hood, 2004; Hood et al., 2015; Hughes et al., 1993;

Mertens, 2007; SenGupta et al., 2004). Although guidance for the use of quantitative approaches are available in the foundational literature for CRE, Chouinard and Cousins (2009) found strong evidence to support the preference of qualitative and mixed-methods approaches. Evaluators using a culturally responsive approach have found that standardized instruments, often normed on dominant, Western and White populations, failed to be culturally relevant (Bowen & Tillman, 2015; Chilisa & Tsheko, 2014; Chouinard & Cram, 2020; Coppens et al., 2006; Pacico et al., 2013) or equivalent (Alkon et al., 2001; Janzen et al., 2015; Sy et al., 2015). Highlighting these challenges in their research, Chouinard and Cousins (2009) questioned the role of quantitative approaches in culturally responsive evaluation and asked "can quantitative approaches help evaluators better engage in culture? If so, which (if any) approaches could be consistent with cross-cultural evaluation (e.g., comparative studies)? Which approaches would further cross-cultural understanding?" (p. 487). This study adds, how are standardized measures identified as culturally responsive?

Measurement across Cultural Groups

The influence of culture in quantitative measures is not unexamined in the measurement literature. Although not often positioned in arguments advancing justice nor equity, measurement research is often concerned with investigations of bias, fairness, and equivalence. The Educational Testing Services (ETS) *Standards for Quality and Fairness* (2016) state, "the most useful definition of fairness for test developers is the extent to which the inferences made on the basis of test scores are valid for different groups of test take takers" (p. 19). When measures fail to be invariant, the data collected may not serve

as an accurate representation of the experiences of the studied population. The most widely used methods for determining equivalence across groups are differential item functioning (DIF) and structural equation modeling. Studies examining DIF used statistical or item response theory techniques, while studies using structural equation modeling often focused on multi-group confirmatory factor analysis.

Measurement research examining invariance across marginalized populations highlights challenges such as establishing increasingly rigorous levels of equivalence across groups (Asil & Brown, 2016; Bryne & van de Vijver, 2010; Bryne & van de Vijver, 2014), cultural bias (Mylonas & Furnham, 2014), and determining why measures and items function differently across groups (Allalouf et al., 1999; Luyt, 2012). Establishing cross-cultural invariance is critical because invariance analysis can signal issues of bias or fairness and serve as evidence of validity. Although research on establishing equivalence across groups is substantial, studies focused on heterogeneous groups or on socioeconomic status across groups are rare and called for (e.g., Asil & Brown, 2016; Oliveri et al., 2016; Pokropek et al., 2017; Randall et al., 2012). Additionally, a number of studies have examined instruments designed for K-12 educational outcomes across demographic groups and across countries (e.g., Abedi et al., 2000; Asil & Brown, 2016; Guttmannova et al., 2008; Kato et al., 2009; Liu & Wilson, 2009), but few have examined outcome attainment in the context of higher education (Lakin et al., 2012).

Validity

Concerns of cultural responsiveness and invariance are concerns of validity. Validity is positioned at the crossroads of measurement, equivalence, culture and cultural responsiveness. Multicultural validity centers validity arguments in culture and is defined as "the accuracy or trustworthiness of understandings and judgements, actions, and consequences, across multiple, intersecting dimensions of cultural diversity" (Kirkhart, 2010, p. 401). Multicultural validity is an argument-based approach to establishing evidence for construct validity originally discussed by Messick (1995) and later, Kane (2013). More broadly, the *Standards for Educational and Psychological Testing* by the American Educational Research Association (AERA), American Psychological Association (APA), and National Council for Measurement in Education (NCME; 1999) define validity as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of the tests…It is the interpretations of test scores required by proposed uses that are evaluated, not the test itself" (p. 9). Validity does not rest with the measure itself, but in data collected across contexts and populations.

In 1995, Messick proposed an approach which unified content, criterion, and construct validity under one theory of construct validity which addressed the meaning, interpretation, and use of scores, and included content, substantive, structural, generalizable, external, and consequential as aspects of validity. More recently, Kane (2013) argued that validity is a product of the explicit statement of both score use and interpretation and proposed an eight-point model for crafting validity arguments. Kirkhart

(2013) builds on Messick's (1995) unified theory of construct validity and centers validity arguments in culture. Multicultural validity outlines five areas of justification or threats to validity including methodological, experiential, relational, theoretical and consequential. Like Messick (1995), Kirkhart positions validity within multiple aspects and extends validity arguments by recommending they attend to history, location, power, voice, relationships, time, reciprocity, plasticity and reflexivity (2013).

The purpose of NSSE is to be a rigorous measure of educationally purposeful activities connected to attainment outcomes which institutions can use to make improvements (NSSE, 2018). Multicultural validity extends validity to ensure that information is accurate and valid across marginalized populations (Kirkhart, 1995). If NSSE is not culturally responsive and invariant for first-generation college students, inferences drawn about these students may be invalid.

Problem Statement

Higher education was historically designed to serve the elite and privileged members of society. Students falling outside of these parameters, like first-generation college students, have been historically underrepresented and underserved. In higher education, there is a gap in our collective understanding of how outcomes data collected with standardized instruments and used to respond to accountability demands, is accurate and trustworthy for first-generation college students. Research in culturally responsive evaluation and measurement on examining invariance across groups highlights that quantitative measures, standardized on dominant populations, lack cultural responsiveness and equivalence (Alkon, et al., 2001; Bowen & Tillman,

2015; Chilisa & Tsheko, 2014; Coppens et al., 2006; Janzen et al., 2015; Pacico, et al., 2013; Sy et al., 2015). Issues of cultural responsiveness and equivalence are issues of validity. Multicultural validity (Kirkhart, 1995) centers validity arguments in culture and privileges marginalized perspectives in order to make valid and fair inferences about the experiences of marginalized groups.

Quantitative measures are often assumed to function equally well for marginalized populations with implications for multicultural validity. Kirkhart (2010) writes, “when it is not visibly identified, the default operating is a dominant majority perspective. Persons with non-majority identification become distanced or treated as 'other,' often with oppressive consequences" (p. 402). Failing to critically examine if a measure is culturally responsive and invariant upholds normative assumptions that all student experiences and knowledge are accurately captured. NSSE is a prolific measure of outcomes in higher education which has gone unexamined for cultural responsiveness and invariance for first-generation college students. How can quantitative data be used to make improvements in higher education, advance attainment outcomes for diverse students, and advance equity when the data itself may not be an accurate representation of diverse student experiences? Examining NSSE for cultural responsiveness challenges the normative assumption that this measure is unbiased and culturally responsive, creating the opportunity to elevate marginalized experiences and advance equity.

Study Purpose and Research Questions

The purpose of this research is to identify and employ strategies to determine to what extent the National Survey of Student Engagement is culturally responsive and

invariant for first-generation college students. Expanding upon existing research, this study extends the work of Chouinard and Cousins (2009) in seeking to understand if there are quantitative approaches which are culturally responsive and further cultural understanding, and it responds to calls in the measurement literature to expand invariance studies to diverse groups such as those with differing socioeconomic statuses. The impetus for this work is in examining approaches to quantitative data collection and analysis in culturally responsive evaluation and measurement to determine to what extent a measure of student outcomes, NSSE, is culturally responsive and provides valid data for a marginalized group, in this context, first-generation college students. Drawing from approaches that privilege culture and position culturally responsive measures in establishing standards of multicultural validity to look at the National Survey of Student Engagement and first-generation college students, this study examines the following research questions:

1. To what extent is the National Survey of Student Engagement a culturally responsive measure for first-generation college students? What considerations for rendering quantitative measures culturally responsive can be derived from a critical examination of the empirical literature on culturally responsive evaluation and measurement?
2. To what extent do statistical techniques used in measurement, such as multi-group confirmatory factor analysis, establish measurement invariance in the National Survey for Student Engagement for first-generation college students? How does this approach further cross-cultural understanding?

Significance

Challenges of accurate measures of diverse, underrepresented students have been cited across two major fields (evaluation and measurement), but with little attention given to the higher education context (Lakin et al., 2012). The primary way in which this research study is significant is that it draws on two well established fields to respond to calls for cultural responsiveness in higher education by aiming to develop critical considerations for evaluating existing measures as culturally responsive and invariant. The outcomes of this study could serve as the building blocks from which to develop a robust set of criteria for evaluating if other standardized surveys are culturally responsive for marginalized populations. Second, this study takes an existing measure and centers first-generation college students in a critical examination of how the survey functions. In the past, the developers of the National Survey of Student Engagement have demonstrated a willingness to change and adapt their survey to strengthen validity claims. Depending on the outcome of this study, the survey could be updated, or additional validity claims could be made for the use of the survey with first-generation college students. Third, the current study adds to the ongoing body of literature in both culturally responsive evaluation and measurement by extending the conversation around measurement equivalence and multicultural validity into higher education, focusing on first-generation college students. Fourth, this study responds to calls in the measurement literature to continue expanding research across socioeconomic groups.

Manuscript Organization

In total, this manuscript is composed of six chapters. Chapter I covered context and a broad overview of and rationale for this study. Chapter II provides a review of the literature related to culturally responsive evaluation, measurement across cultures, validity and multicultural validity. Chapter III covers an overview of the methodology for a critical examination of the empirical literature and establishing invariance in the National Survey of Student Engagement. Chapters IV and V, provide results and study specific discussions. Chapter VI includes a broader discussion, limitations, suggestions for future research, and conclusions.

CHAPTER II

LITERATURE REVIEW

In the context of higher education, an increasingly diverse student body navigates academic settings, co-curricular engagement opportunities, and environments that were historically designed for affluent White men. While institutions have evolved and adapted to better meet the needs of marginalized students and at the same time respond to accountability demands, differential outcomes persist across groups. What remains unclear is whether standardized surveys, such as the National Survey of Student Engagement, used to measure student outcomes, are culturally responsive and invariant measures which provide accurate and trustworthy data reflective of historically underrepresented student experiences. Failing to interrogate if standardized surveys are culturally responsive and accurately measure outcomes for a diverse student body privileges the dominant experience, effectively erasing experiences considered "outside" this assumption. Questioning the efficacy of these instruments challenges researchers to examine the status quo and creates space for multiple realities and lived experiences. This study seeks to further cross-cultural understanding in a way that balances demands for scientific rigor, while privileging historically marginalized experiences in the context of higher education. An examination of foundational and empirical literature in culturally responsive evaluation, measurement equivalence across cultural groups, and multicultural validity is needed in order to interrogate cultural responsiveness and invariance in

quantitative measures. This study uses the National Survey of Student Engagement and first-generation college students as a case example to examine cultural responsiveness and invariance in quantitative measures; therefore, an explanation of the survey and first-generation college students as a cultural population are also included.

Culturally Responsive Evaluation

Critical to understanding culturally responsive evaluation are concepts of cultural responsiveness, the current landscape of CRE and quantitative methods, and guidance for the use of quantitative methodologies and methods provided in CRE literature. In this section, each of these topics are explored in detail to provide a theoretical lens for examining cultural responsiveness in quantitative measures.

Cultural Responsiveness

Culture is a shared set of values and behaviors within a group of individuals, which can be characterized by demographic variables and systemic factors such as politics and economics. At the individual level, culture is characterized by demographic variables, worldviews, and behaviors. Culture is fluid and shapes beliefs and worldviews (SenGupta et al., 2004), ways of meaning making and knowing (AEA, 2011; Bocock, 1992; Nastasi & Hitchcock, 2016), and behaviors (Frierson et al., 2002; Guzmán, 2003; Nastasi & Hitchcock, 2016). At the systemic level, culture influences context, economics, and politics (AEA, 2011). All individuals develop in contexts shaped by culture, which influences what is taught, how learning occurs, how learning is demonstrated, which ways of learning are considered valid, and the tools used to measure learning (Hughes et al., 1993).

In evaluation practice, “to be responsive does not automatically yield design authority to stakeholders. It means coming to know the circumstances and problems and values well, then using professional talent and discipline to carry out the inquiry” (Abma and Stake, 2001, p. 9). Responsive approaches to evaluation are interactive, reflective, and participatory, framing stakeholder needs and concerns as central to the evaluation process, while evaluators still lead the evaluation design and implementation (Hopson, 2009). Cultural responsiveness in evaluation has theoretical roots stemming from indigenous and minoritized ways of knowing and meaning making, democratic principles, social change, feminist and critical theory (Hopson, 2009). Culturally responsive evaluation expands on responsive evaluation by attending to issues of power, race, equity, culturally related contextual factors such as demographics and socioeconomic factors, and social justice (SenGupta, et al., 2004). Cultural responsiveness brings historically marginalized populations in from the margins, centering minoritized ways of knowing and meaning making throughout the evaluation process to reduce further marginalization and the reproduction of injustices (Bledsoe & Donaldson, 2015; Mertens, 2010).

Situating Cultural Responsiveness in Evaluation

Broadly, evaluation is a practice driven by asking and answering questions in a specified context to determine the value, merit, or worth of a program (Scriven, 1991). More specifically, Patton (2008) defined evaluation as, “the systematic collection of information about the activities, characteristics, and results of programs to make judgments about the program, improve or further develop program effectiveness, inform

decisions about future programming, and/or increase understanding" (p. 37). Evaluation extends to a variety of contexts to answer a range of formative and summative questions, and is grounded in social science reflective of a history rooted in power and colonization (Stanfield, 1999). Comparisons of 'us' and 'other' across cultures stem from roots in European expansion and colonialism, and often center and validate White ways of knowing (Hall, 1992). Theses comparisons have led to a developed discourse around the "West" and "the rest" (where West became synonymous with civilized and the rest with uncivilized people), which still shapes modern notions of scientific investigation and what counts as valid knowledge (Hall, 1992; Hood et al., 2015; Hopson, 2009). One of the lasting effects of this history is that majority, dominant, perspectives continue to shape which questions are prioritized and how they are measured, creating an ethnocentric approach to research, maintaining existing power structures, and perpetuating inequalities and injustices (Gordon, Miller, & Rollock, 1990; Hughes, et al., 1993; Stanfield, 1999).

Different approaches to evaluation prioritize the use of results, methodology, value judgements, and justice. Christie and Alkin (2013) provide a metaphor for describing evaluation theory as a tree with roots in social accountability, social inquiry, and epistemology. Stemming from these roots, are three branches of evaluation grounded in use (how and who will use the evaluation), methods (research methodology), and valuing (the subject and context of the evaluation) (Christie & Alkin, 2013). Over the last thirty years, there has been an additional focus on culture and its influence on the evaluation process (Hood et al., 2005; 2015) and an emergence of new approaches such

as transformative evaluation (Mertens, 2007) and culturally responsive evaluation (Frierson et al., 2002). Mertens and Wilson (2012) revised the tree, adding a social justice branch to capture transformative and culturally responsive approaches. Transformative evaluation focuses on challenging power structures rooted in oppression, building trust, engaging with the community, and sharing findings to advance human rights and justice (Mertens, 2007; 2009). Culturally responsive evaluation, a transformative approach, seeks to unearth historical context, redistribute power, and center culture, responsiveness, and context in the evaluation process (Frierson et al., 2002; Hood et al., 2015).

Culturally Responsive Evaluation

To redistribute power and unearth historical context, culturally responsive evaluators reject the idea of evaluation as culture free or value neutral (Frierson et al., 2002; Hood et al., 2015; Hopson, 2009). Instead, culturally responsive evaluation integrates culture and the lived experiences of those participating in the evaluation throughout the evaluation process in preparing the evaluation, developing questions, selecting and engaging stakeholders, identifying the purpose of the evaluation, selecting or developing methods, and collecting, analyzing and reporting the data (Frierson et al., 2002). Culturally responsive evaluation is not a new approach, rather it is an approach to evaluation conducted to "create accurate, valid, and culturally-grounded understandings" (Hood et al., 2015, p. 291). Culturally responsive evaluation centers culture in the evaluation while also attending to context and responsiveness (Greene, 2015). Context includes the descriptive and demographic characteristics, material and economic features, institutional and organizational climate, interpersonal dimensions, and the political

dimensions of a setting (Greene, 2005; SenGupta et al., 2004). Context, beyond the physical location of the evaluation, includes shaping how the evaluation unfolds, establishing trust with stakeholders, and navigating power dynamics.

Context, cultures, and approaches to responsiveness may be unique across evaluations, but evaluator reflexivity, centering cultural perspectives, addressing power, advancing justice, and multiple methods are foundational to the practice of CRE. Evaluator *reflexivity* involves understanding one's own cultural location and biases and how cultural lens shapes the evaluation process and subsequent findings (AEA, 2011; Gordon et al., 1990; Hopson, 2003; SenGupta et al., 2004; Symonette, 2004). The American Evaluation Association (AEA) statement on cultural competence links competence and reflexivity. Reflexivity fosters cultural competence to the extent that evaluators actively and persistently engage in self-examination to readily identify and attend to the ways in which their lived experiences and backgrounds may serve to strengthen or impair an evaluation (AEA, 2011).

Culturally responsive evaluation approaches *embrace cultural perspectives* through stakeholder involvement and centering stakeholder lived experiences as expertise (Frierson et al., 2002; Hood, et al., 2015; Hopson, 2009; Kirkhart, 1995). As Hopson (2009) explains,

> Those who use CRE understand and value lived experiences that help to (re)define, (re)interpret, and make sense in everyday life. By privileging notions of lived experiences and especially regarding communities and populations of color or indigenous groups, new explanations and understandings of evaluands, programs, and phenomena of study emerge. (p. 431)

Integrating stakeholder perspectives allows for more meaningful conceptualizations of culture beyond demographic variables, serves as a means to unearth inequities, and creates space for multiple realities. The role of evaluators in *advancing social justice* and *explicitly addressing power* dynamics in the context of the evaluation is tied to evaluator positionality and stakeholder integration in the evaluation process (Hood et al., 2015; Hopson, 2003; Hopson, 2009). Evaluators must be aware of their social locations and have the cultural competence to know when they may be disrupting or upholding systems of oppression. Engaging stakeholders is a common avenue for redistributing power, challenging bias, and avoiding ethnocentric approaches to evaluation. Evaluator positionality and stakeholder culture shapes choices related to *methodology and methods* of data collection (AEA, 2011; Chouinard, 2016; Frierson et al., 2002). To attend to these influences, evaluators frequently engage in the use of multiple methodologies and methods of data collection (Chouinard and Cousins, 2009; Frierson, et al., 2002; Hood, et al., 2015; Hughes et al., 1993). Methodology and methods are critical considerations in CRE as choice of method impacts how culture is represented, conceptualizations of voice and knowledge creation, and epistemological advancements.

Culturally Responsive Evaluation and Quantitative Approaches

Methodologies and methods are values-based, shaped by context and culture, influence the validity and credibility of data, and require decisions shaped by evaluator positionality all of which have implications for equity and justice (AEA, 2011; Frierson et al., 2002; Hood, 2004; Hood et al., 2015; Hughes et al., 1993; Mertens, 2007; SenGupta et al., 2004). Considering culture in the selection of methodologies and

methods is critical as these choices shape who is heard and how knowledge is shaped. Dominant perspectives are reinforced or disrupted through methodological choices. Gordon et al. (1990), argue that such choices reflect the "communicentric bias" of the researcher and suggests adopting approaches to inquiry that better accommodate cultural populations. Culture intersects with methodology and methods through epistemology, instrument development, selection and adaptation, levels of inclusion or exclusion of voice, definitions of credible evidence, and multicultural validity (Chouinard, 2016). Culturally responsive evaluators must respond to cultural context as well as identify and examine any underlying assumptions present in the selection or construction, implementation, and analysis of methods (Hood et al., 2015; Hughes et al., 1993). Culture shapes what we learn, how we learn, and how we demonstrate learning. Culturally responsive approaches attend to cultural nuances in the evaluation process. Building on the core aspects of culturally responsive theory and practice discussed above, the rest of this section explores the landscape of culturally responsive evaluation and the use of quantitative methodologies and methods. Critical to examining this landscape are discussions of methodology and methods, standardized tools used across cultural populations, the tension between cultural responsiveness and generalizability, and recommendations for the use of quantitative methods from culturally responsive evaluation scholars.

Methodology is defined as the guiding philosophy shaping method selection (Carter & Little, 2007); methods are the mechanisms (e.g., surveys, case studies, focus groups, and photo voice) used to collect data which can take the form of narratives,

pictures, and numbers. Programs cannot be accurately measured and understood without attending to the cultural context which influences the design of the programs or the tools used to measure them (Bledsoe & Donaldson, 2015; Frierson et al., 2010; Hood et al., 2005; SenGupta et al., 2004). Epistemology, or knowledge creation, is tied to choices of methodology and methods (Carter & Little, 2007; Chouinard & Cousins, 2009). Questions of epistemology draw attention to how evaluators create the space for stakeholders to be experts in their own knowledge construction, sharing, and demonstration. Hughes et al. (1993) write that culture permeates the context in which humans grow and develop, and as such, “this context supplies blueprints for living that determine what is learned, the process through which learning occurs, and the rules for displaying competencies that are valued by group members" (p. 689). Methodology is the guiding philosophy which shapes choices of method for data collection; methods shape what type of data is collected, how voices are heard and represented, shaping knowledge construction.

Prior research in CRE noted that evaluators using culturally responsive approaches often used qualitative and mixed methods, drawing attention to the lack of quantitative instruments designed and normed for cultural groups (Chouinard & Cousins, 2009; Frierson et al., 2002). The use of multiple methodologies and methods of data collection allows evaluators to explore and address diverse and complex cultural communities in which programs occur (Frierson, Hood, Hughes, & Thomas, 2010; Hood et al., 2015). Methodological pluralism fosters the ability of evaluators to tell a comprehensive story regarding the program, mitigate power dynamics (e.g., Christie and

Barela, 2005), and better develop methods of data collection for valid results (Butty et al., 2004; Coppens et al., 2005). Evaluators using a culturally responsive approach have found that standardized instruments, often normed on dominant populations then used in marginalized populations, failed to be culturally relevant or equivalent (Alkon et al., 2001; Bowen & Tillman, 2015; Chilisa & Tsheko, 2014; Coppens et al., 2006; Janzen et al., 2015; Pacico et al., 2013; Sy et al., 2015). In a review of 52 empirical articles focused on culturally responsive evaluations, authors Chouinard and Cousins (2009) found that the use of standardized instruments, which had not been validated for diverse populations, created significant challenges. They subsequently posed questions regarding measurement and conceptual equivalence of quantitative tools across populations and contexts. The same challenge has persisted over the last decade (e.g., Bowmen & Tillman, 2015; Janzen et al., 2015; Sy et al., 2015). Additionally, Frierson, Hood, and Hughes (2002) cautioned that instruments normed on dominant populations may have established validity and reliability, but are not culturally responsive, and may be inappropriate or irrelevant for a given cultural population. The lack of available instruments normed for marginalized populations prevents evaluators from drawing valid inferences about the needs and experiences of marginalized populations. When the measures used to collect and analyze data with the intention of drawing conclusions about program worth or to make large generalization are not critically examined for responsiveness and bias, they can serve to further marginalize and silence cultural experiences and uphold oppressive structures, or the status quo.

As an example, Alkon et al. (2001) experienced a number of challenges at the intersection of culture and methods for data collection when evaluating a violence prevention program serving ethnically diverse children and families in a child care environment. In this study, the evaluators used survey instruments/questionnaires, focus groups, observations and interviews to measure the impact of the program and progress towards outcome achievement. A major consideration was creating instruments and items that were culturally relevant and equivalent. Alkon et al. (2001) spoke to the importance of equivalence across cultural groups writing, "most relevant standardized instruments had been validated with one ethnic group, usually European Americans, and did not have information on conceptual equivalence for other ethnic groups" (p. 51). They continued to explain why this is problematic, "administering instruments to participants from different ethnic groups can be a problem if no linguistic, conceptual, or measurement equivalence is established because it is then difficult to interpret findings for these different ethnic groups" (Alkon et al., 2001, p. 51-52). For measures not validated in different ethnic populations, the evaluators pretested items, made modifications, and when discrepancies existed across groups, focus groups were conducted (Alkon et al, 2001). In addition to the importance of establishing equivalence, Alkon et al. (2001) emphasized the critical role of cultural perspectives in the development and adaption of quantitative tools. In reflecting on their experience working with multiple cultural groups using multiple methods, they discussed the importance of understanding the cultural meaning of items, obtaining demographic information around cultural values in addition

to identities, and establishing validity evidence for each quantitative measure used with each population.

In a more recent example, Bowen and Tillman (2015), took a culturally responsive approach to the design, implementation, and analysis of three surveys used to measure the experiences of a former fugitive slave community (*quilombos)* in Brazil. They conducted extensive field research and focus groups in the development of their surveys. Stakeholders in this evaluation played a significant role in providing feedback on the surveys. In their effort to take a culturally responsive approach to quantitative data collection, the authors reflected that, "culturally responsive inquirers need to acknowledge and address the potential tension between conventional methods of quantitative instrument development, data collection, and analysis, and the desire to be CRE centered" (Bowen & Tillman, 2015, p. 38). They explain, "while we needed to collect standardized measurements, the World Bank survey instrument was not sensitive to the cultural aspects of the local community or the general socioeconomic structure of Brazil" (Bowen & Tillman, 2015, p. 38). Where Alkon et al. (2001) argued for the necessity of surveys which are culturally responsive, Bowen and Tillman (2015) actively engaged in the work of developing, implementing and analyzing the results of a culturally responsive survey and discussed the challenges inherent in that work.

These two examples illustrate a significant tension between broad generalizability and cultural responsiveness in quantitative measures. In working to evaluate an assistance program for Indigenous community members, Martinez et al. (2018) wrote, "in addressing questions that are important for a western scientific audience, evaluators

invariably overlook more relevant and valid areas of cultural learning and development" (p. 33). Speaking more broadly to this same tension, Hughes et al. (1993) argue:

> The culturally anchored researcher must weigh the trade-offs between sensitivity to cultural nuances of the target population and the methodological requirements of objectivity, standardization, and generalizability. In this way, we can begin to develop a knowledge base of diverse cultural groups that balances the demands for rigor and sensitivity. (p. 700)

Central to this tension is context. Hughes et al. (1993) acknowledge this tension, but deny its resolution as a dichotomous choice, instead suggesting a middle ground may be possible to address this tension through the use of new research approaches. Garaway (1996) wrote the "evaluation required construction of an approach amenable to a cross-cultural analysis; in other words, an approach that would produce some over-arching answers, yet still remain sensitive and flexible enough to portray the innuendos and idiosyncrasies of any one given location" (p. 203). For data to be standardized and generalizable, context is often removed as much as possible. For measures to be culturally responsive, they must often balance standardization and comparability of the data with the relevance and credibility in the cultural context (Sutton et al., 2016; Uhl et al, 2004).

Given this tension between local and global, specific and generalizable, identifying how culturally responsive evaluators recommend employing quantitative methods of data collection and analysis is critical. Recommendations for the use of quantitative approaches to data collection and analysis stemming from the foundational literature on CRE include engaging in evaluator reflexivity and representation of voice,

establishing equivalence, considering contextual conditions in data analysis, conducting analysis between and within groups, and engaging multiple forms of analysis. Evaluator *positionality* shapes what evaluators see and hear in evidence collection and also shapes data interpretation and reporting. Culturally responsive evaluators engage in critical self-reflection to appreciate how their positionality shapes their work (Hood et al., 2015) and work to ensure cultural experiences emerge in the data interpretation and presentation (Frierson et al., 2010). Like Alkon et al. (2001), culturally responsive evaluators using quantitative measures seek to establish *equivalence and validity* evidence in new context and for cultural populations (AEA, 2011; Hood et al., 2015; Hughes et al., 1993). Hughes et al. (1993) argue that even between group studies are only valid when, "the concepts, measures, and tasks are equivalent across groups. A culturally anchored methodologist must examine issues of equivalence in depth" (p. 695). Examining equivalence ensures the same constructs are measured across groups and provides evidence of validity.

Related to voice and reflexivity is data analysis and dissemination, scholars in CRE recommend that evaluators remain aware of the frame through which they interpret and analyze data, use *multiple approaches to examine data*, and focus on both *between and within group* studies. Frierson et al. (2010) write that,

> Data do not speak for themselves nor are they self-evident; rather, they are given voice by those who interpret them. The voices that are heard are not only those who are participating in the project, but also those of the analysts who are interpreting and presenting the data. (p. 91)

When evaluators practice reflexivity and develop cultural competence, they can better recognize how their bias and lived experiences shape data analysis and interpretation.

Engaging with stakeholders and using multiple methods can also mitigate evaluator bias. Hood et al. (2015) recommend taking an investigative approach to data analysis by considering intended and unintended outcomes emergent in the data (Hood et al., 2015), disaggregating data (Hood et al., 2015; Frierson et al., 2010), and conducting within and between group analysis (Gordon et al., 1990; Frierson et al., 2010; Hood et al., 2015; Hughes et al., 1993). Frierson et al., (2010) argue that "culturally responsive evaluations use multiple strategies to analyze quantitative data to reveal a more complete picture of what is occurring within the environment under study" (p. 90) which serves to contextualize the data. An example of decontextualized data from higher education would be drawing conclusions from a standardized survey about the attainment outcomes of marginalized students without considering the contextual factors influencing their attainment. Decontextualized outcomes data interpreted by culturally incompetent practitioners positions attainment as a singular outcome unaffected by student experiences or the surrounding environment and reinforces the faulty normative assumptions that all students have the same experience. Finally, when conducting quantitative analysis and cross-group comparisons, culturally responsive evaluators work to avoid any deficit based language or conclusions (AEA, 2011; Hughes et al., 1993) and avoid comparisons which affirm "Whiteness" as the standard to which all group are compared (AEA, 2011).

Given the challenges with standardized instruments, in summarizing their directions for future research, Chouinard and Cousins (2009) asked, "can quantitative approaches help evaluators better engage culture? If so, which (if any) approaches would

be consistent with cross-cultural evaluation (e.g., comparative studies)? Which approaches would further cross-cultural understanding?" (Chouinard & Cousins, 2009, p. 31). The challenge of standardized measures in CRE has been well documented, but the ways in which culturally responsive evaluators determine a measure to be culturally responsive is largely unexamined in the literature. In addition to asking if quantitative measures can advance cross-cultural understanding, this study asks how culturally responsive evaluators interrogate quantitative measures for cultural responsiveness and invariance.

Measurement

Equivalence across diverse populations is a concern repeatedly noted for evaluators using a culturally responsive approach. In an effort to attend to historical context and center cultural perspectives, culturally responsive evaluators using quantitative measures must interrogate assumptions that such measures collect valid, accurate, and trustworthy evidence across cultural contexts and populations (AEA, 2011; Hood et al., 2015). Drawing on the measurement literature, a more detailed examination of equivalence, how it is related to culture, why it is important, and how it is established are addressed in this section.

Measurement is concerned with systematically assigning numbers to people in order to represent properties an individual might have (Allen & Yen, 1979, p. 2), such as knowledge around math or science, skill development, HIV awareness, or violence prevention strategies. Measurement theory, or the study of measurement, is "a branch of applied statistics that attempts to describe, categorize, and evaluate the quality of

measurements, improve the usefulness, accuracy, and meaningfulness of measurements, and propose methods for developing new and better measurement instruments" (Allen & Yen, 1979, p. 2). To achieve these ends, a multitude of studies examine the interaction of student demographics with tests items. In measurement, evidence of trustworthy data is established, in part, by examining how tests and surveys function across groups. As such, examining how tests function across diverse groups and cultures emerges in the measurement literature through analyzing tests and items for differential functioning (e.g., Elosua & López-jaúregui, 2007; Oliveri et al., 2016), test translation and adaptation (e.g., Hambleton, Merenda, & Spielberger, 2005), and equivalence across cultures (e.g., Asil & Brown, 2016; Avery et al., 2007; Lakin et al., 2012).

Equivalence, also referred to as measurement invariance, “concerns whether scores from the operationalization of a construct have the same meaning under different conditions” which include “consistency of measurement over populations, time of measurement, or method of test administration” (Kline, 2011, p. 251). Hughes et al. (1993) argued that the relevance of constructs measured varies by cultural population and “the construct as well as the range of relevant items that adequately assess the construct may be very different for members of different populations” (p. 692). Additionally, behavior can have different meanings, relationships between indicators and constructs may differ, and the attitudes/values relevant to a construct all may differ across cultural populations (Hughes et al., 1993). Establishing invariance provides evidence for validity in that the same constructs are measured across groups.

Examining and establishing invariance across cultural groups serves as evidence of data validity and lack of bias in an instrument. When evidence of invariance is established, this provides evidence that a measure functions similarly across groups, allowing for cross-group comparisons and supporting decision making with the data. Concepts of bias, fairness, and equivalence are inextricably linked (Penfield, 2016; AERA, APA, & NCME, 2014). When items do not function similarly across groups, the ability to draw valid inferences across marginalized populations is inhibited and may lead to inaccurate conclusions about the construct of study across groups (Zumbo, 1999). For example, in higher education, such conclusions may suggest engagement outcomes vary across groups and while this may be true due to item impact, item bias and construct irrelevant variance must be examined to be sure. In the organization's statement on cultural competence in evaluation the American Evaluation Association (2011) wrote, "inaccurate or incomplete understandings of culture introduce systematic error that threatens validity. Culturally competent evaluators work to minimize error grounded in cultural biases, stereotypes, and lack of shared worldviews among stakeholders" (p. 5). Test items are designed to solicit item impact, item impact happens when an item accurately differentiates between groups, determining true differences between those with a given trait and those without. Items which give one group an unfair advantage, based on the underlying trait measured by the item are said to be biased (Zumbo, 1999). Bias occurs when items, constructs, or tests function differently across groups and is often studied through invariance testing. Differential item functioning (DIF) analysis and

confirmatory factor analysis are two of the most widely used approaches for establishing invariance and examining item and test bias.

Differential Item Functioning

Differential item functioning analyses are used to examine systematic test and item bias across groups. Differential item functioning happens when although two students (e.g., one female and one genderqueer) have the same total test score or estimated ability level, they perform differently on a given item as a result of some property of the item (Penfield, 2016). In other words, the item is measuring the construct as well as a secondary construct (in this case, gender). An item which displays DIF works in favor of one group and against another. Differential item functioning is not synonymous with item bias but can serve as an indicator of bias with additional analysis (Penfield, 2016). Researchers use a collection of statistical and item response theory approaches for identifying if dichotomous and polytomous items function differently across groups after conditioning for ability. These approaches extend statistical analysis beyond a comparison of means in order to identify if constructs unrelated to those measured on the assessment (e.g. group membership) influence (e.g. favor or disadvantage) one group over another. Examinations of DIF do not explain why an item or test may function differently across groups, but the different approaches to calculating DIF do identify if there is bias and in some cases, the severity of that bias. Prior to conducting DIF analysis, many studies examine the internal structure of the data in order to identify issues related to factors or item and factor relationships (e.g., Lakin et al., 2012; Randall et al., 2012). Item response theory approaches are largely used when a

single construct is measured, where confirmatory factor analysis can be used to examine the internal structure of the data, and when multiple underlying constructs are measured as is the case with NSSE.

Multi-Group Confirmatory Factor Analysis

Confirmatory factor analysis techniques fit a specified model to two or more independent samples of data and apply a series of increasingly restrictive criteria to test different levels of equivalence (Byrne et al., 1989; Meredith, 1993). Confirmatory factor analysis (CFA) "analyzes a priori measurement models in which both the number of factors and their correspondence with the indicators are explicitly specified" (Kline, 2011, p. 112). A multi-group confirmatory factor analysis (MG-CFA) is an approach used to test measurement invariance where a measurement model is fitted to at least two independent samples of data (Kline, 2011). Multi-group confirmatory factor analysis is used to examine increasingly restrictive levels of equivalence starting with configural invariance (equal factor structures), metric invariance (factor loadings), and scalar invariance (thresholds). Testing for each of these levels of equivalence is hierarchical, meaning configural must be established before examining metric, which must be established before establishing scalar. Examining equivalence focuses on measurement and structural issues. Measurement equivalence concerns regression intercepts, regression slopes, and error variances that are invariant across groups while structural equivalence concerns invariant factor means and variance-covariance structures (Byrne et al., 1989).

Configural invariance tests to see if both the number of factors and the items loading on each factor are the same across groups (Byrne et al., 1989; Cheung &

Rensvold, 2002). If the same model fits each group, then the same underlying, or latent, constructs are measured, and equivalence at this level holds. Configural invariance specifies and fits the same measurement model to two separate groups by keeping the number of factors and their associated indicators the same, while freely estimating all parameters (Byrne et al., 1989; Meredith, 1993). Configural invariance keeps the same pattern of factor loadings while allowing the weight of the factor loadings, the thresholds, and the error variances to be freely estimated. In establishing configural invariance, the same pattern of loadings should occur, but the weight of these loadings can differ as well as the intercepts and error variances, at this point the researcher can conclude that the same constructs are present for both groups (Kline, 2011).

The next level is metric invariance which tests if the factor loadings for each item are equal across groups (Kline, 2011). Metric level invariance shows that not only are the same constructs measured across groups, but the constructs "manifest the same way in each group" (Kline, 2011, p. 253). To test the metric model, factor loadings are constrained for each indicator across groups and the error variance and thresholds are freely estimated (Byrne et al., 1989; Cheung & Rensvold, 2002). In this model, the loadings are fixed to be the same, but one group may have more measurement error than the other. Scalar invariance examines item intercepts or thresholds (Byrne et al., 1989; Cheung & Rensvold, 2002). Scalar invariance tests to see if the groups in the MG-CFA use the scale in the same way (Cambell et al., 2008). In the scalar model, both factor loadings and intercepts are fixed. The test for scalar invariance determines if the test performs the same across groups and allows for the comparison of latent means.

Establishing invariance provides a measure of whether or not an instrument functions the same across groups. Using multi-group confirmatory factor analysis to examine invariance is beneficial because this analysis allows for an understanding of test functioning as well as if some items, or constructs, are more relevant to one cultural group or another. Multiple studies in the measurement literature have examined invariance across countries (e.g., Bryne & van de Vijver, 2010; Huang et al., 2016) and within countries across racial groups and gender binary groups (e.g., Avery et al., 2007; Banks, 2006; Bank, 2012; Liu & Wilson, 2009; Wu, 2010); however, studies examining invariance across socioeconomic groups are rare (e.g., Asil & Brown, 2016; Randall et al, 2012) and called for (e.g., Ayalon & Young, 2009). The *Standards for Educational and Psychological Testing* (1999) encourage expanding the examination of outcomes across subgroups beyond race and ethnicity as a mechanism for demonstrating fairness and reducing bias in testing. This study responds to this call, using multi-group confirmatory factor analysis as, to date, no studies have applied this analysis to the study of first-generation college students using data from NSSE. Without establishing equivalence, the assumption is that NSSE functions the same across groups and means can be compared in order to make decisions to improve student experiences. If metric and scalar equivalence are not established, comparisons across groups are inappropriate.

Validity

Positioned at the intersection of measurement, equivalence, culture and cultural responsiveness, are concerns of validity. Related to validity is the need to draw accurate inferences across diverse populations, further culturally grounded understandings,

improve the accuracy of findings, and minimize conclusions shaped by stereotyping and researcher bias. Critical to understanding the role of validity in this study are concepts of how validity is defined, including multicultural validity, and establishing justifications and arguments for validity. Challenges to cultural responsiveness and invariance are issues of validity. According to the *Standards for Educational and Psychological Testing* (1999):

> Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of the tests. Validity is, therefore, the most fundamental considerations in developing and evaluating tests. The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores required by propose uses that are evaluated, not the test itself. (AERA, APA, & NCME, p. 9)

Messick (1995) presented a theory of validity which bridged together the existing forms of criterion, content, and construct validity under one unified theory of construct validity. He wrote, "construct validity is based on an integration of any evidence that bears on the interpretation or meaning of the test scores – including content- and criterion-related evidence" (Messick, 1995, p. 742). More recently, Kane (2013) argued that validity is a product of the explicit statement of both score use and interpretation and proposed an eight-point model for crafting validity arguments. Messick (1995) presented validity evidence as an ongoing accumulation of evidence that amasses to present a case for data collected for a specific assessment. In contrast, Kane (2013) creates boundaries for the development of validity evidence by first creating an argument for validity within a specific context and then providing evidence to support that argument. Kane (2013)

refers to this as the interpretation/use argument (IUA). The agreement should provide enough detail so as to serve as an overarching framework for reporting critical evidence to support data interpretations and uses (Kane, 2013). He organizes and prioritizes validity evidence in four areas of scoring, generalization, extrapolation, and implication with recommendations for collecting supporting evidence to validate the inferences drawn at each stage (Kane, 2013). Scoring inferences draw from observed performances on a test to an observed score, while generalization inferences draw from a sample of performances to performance on a larger domain, extrapolation inferences are used to predict future scores, and implication inferences are related to the intended and unintended consequences of score interpretation and use for decision making (Kane, 2013).

Multicultural Validity

Centering culture in an arguments-based approach to validity is multicultural validity. Multicultural validity is not a new form of validity; rather it is about attending to culture when developing validity evidence. According to Kirkhart (2010), “multicultural validity refers to the accuracy or trustworthiness of understandings and judgements, actions, and consequences, across multiple, intersecting dimensions of cultural diversity" (p. 401). She argues for a more inclusive definition of validity as “an overall judgement of the adequacy and appropriateness of evaluation-based inferences and actions and their respective consequences” (Kirkhart, 2005, p. 30). In this definition, validity is defined in a way that addresses multiple ways of knowledge construction, decisions made with the data and subsequent consequences, and the social justice implications (Kirkhart, 2005).

Multicultural validity includes measurement validity, interpersonal validity, and consequential validity (Kirkhart, 1995). Measurement validity is concerned with the use of tools which are both relevant to the life experiences of the people responding and equivalent across groups. Interpersonal validity is concerned with personal interactions in the data collection process. Consequential validity concerns the actions resulting from an evaluation, both negative and positive, as well as intended and unintended consequences. The purpose of NSSE is to be a rigorous measure of educationally purposeful activities connected to attainment outcomes, which institutions can use to make improvements (NSSE, 2018). Multicultural validity extends to ensuring that information is accurate and valid across stakeholder groups, especially marginalized populations, and encompasses measurement validity, design logic validity, interpersonal validity and consequential validity (Kirkhart, 1995). Kirkhart (2013) outlines five areas of justification or threats to validity including methodological, experiential, relational, theoretical and consequential.

Table *2* outlines the justifications and threats related to each of these five areas of validity evidence.

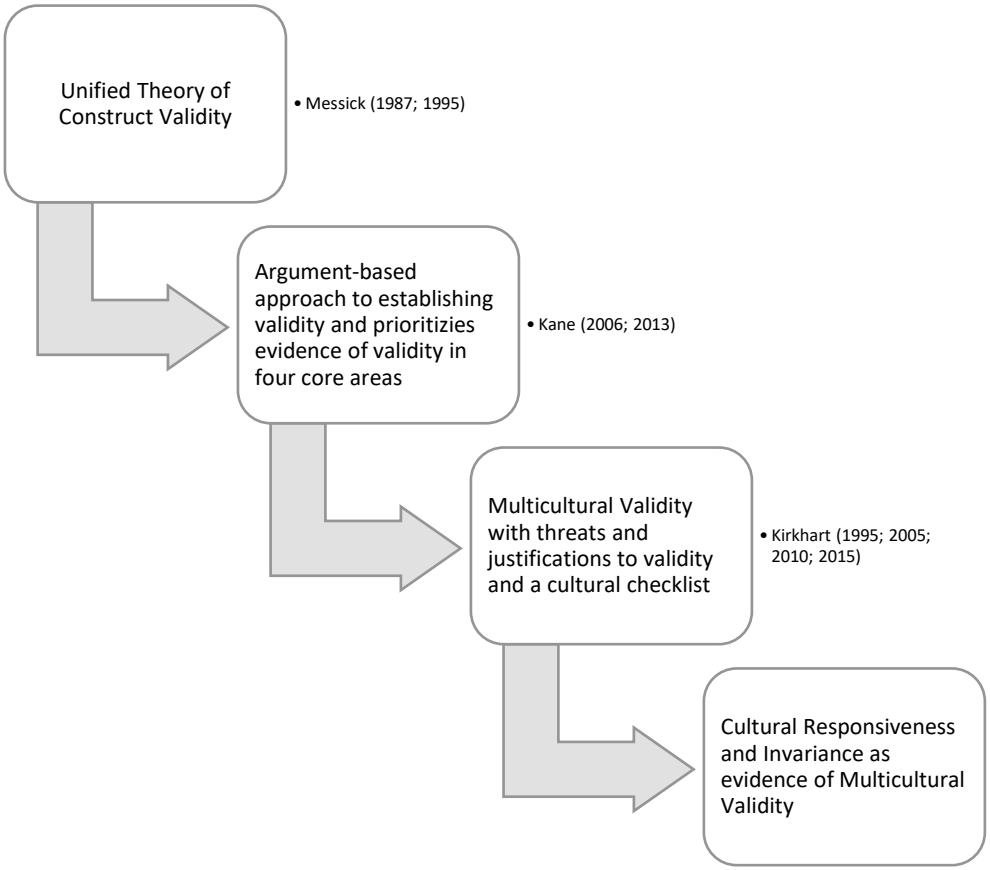
Table 2: Threats and Justifications for Multicultural Validity (Kirkhart, 2013)

	Justifications	**Threats**
Methodological	Attend to choices of methodology, method, and epistemology	Choices of framework, non-responsive methods and procedures, language non-equivalence, imposition of dominant values
Experiential	Include the life experiences of stakeholders	Minimizing, excluding, cultural incompetence, lack of self-reflexivity
Relational	Relationship-oriented, trust and respect	Unapproachable behaviors, lack of honesty, distrust

Theoretical	Scrutinize theoretical justifications	Imposing inappropriate theoretical perspectives, theory incongruent with contexts
Consequential	Consider intended and unintended implications of an evaluation, promote equity and justice, reciprocal, address oppressive histories	Lack valuable information on impact when consequences are unconsidered

In attending to the threats and justifications, Kirkhart (2013) also provides a checklist which considers history, location, power, voice, relationships, time, reciprocity, plasticity, and reflexivity in the argument for validity while moving through an evaluation process. The current study draws on the threats, justifications, and checklist provided by Kirkhart (2013) to establish cultural responsiveness and equivalence as aspects of multicultural validity specifically for first-generation college students and the National Survey of Student Engagement. Figure 1 shows the interconnectedness of the work of Messick (1995), Kane (2013), and Kirkhart (1995; 2005; 2010; 2013), which guided the development of evidence of multicultural validity in questioning if the constructs measured on NSSE are invariant for first-generation college students and if the tool itself is culturally responsive.

Figure 1: Cultural Responsiveness and Invariance as Evidence of Multicultural Validity



Connecting multicultural and validity work brings concerns of cultural populations into the heart of evaluation rigor, embraces cultural ways of knowing, and addresses issues of dominant ways of knowing and meaning making (Kirkhart, 2013). Kirkhart (2010) explained, "when it is not visibly identified, the default operating is a dominant majority perspective. Persons with non-majority identifications become distanced or treated as 'other', often with oppressive consequences" (p. 402). Examining quantitative tools for responsiveness and equivalence across groups challenges this default operating assumption. Related to Kirkhart's explanation, the *Standards for*

Educational and Psychological Testing (1999) standard 7.1 states, "when credible research reports that test scores differ in meaning across examinee subgroups for the type of test in question…the same forms of validity evidence collected for the examinee population as a whole should also be collected for each relevant subgroup" (AERA, APA, & NCME, p. 80). Score meanings may differ across populations and different groups may understand constructs differently. As such, evidence should be collected to verify that subgroups understand constructs similarly and cross-group scores can be compared. Collecting such evidence is congruent with approaches in CRE where "possible interpretations and ways to establish validity of results are proactively discussed, along with how the interpretations will occur and how they will be shaped to be responsive to the needs of the community" (Mertens & Wilson, 2015, p. 200). Establishing invariance across subgroups is evidence of construct validity and suggests that each subgroup understands the constructs measured in similar ways.

A Case Example

Cultural responsiveness and invariance provide evidence for multicultural validity, or accurate and trustworthy evidence for cultural populations. The purpose of this study is to examine how standardized measures can be determined as culturally responsive for marginalized populations using the National Survey of Student Engagement and first-generation college students as a case study. This section provides a history of NSSE, an overview of core concepts measured by NSSE, and research conducted on NSSE. This section also covers first-generation college students as a cultural population historically marginalized in higher education. In this case example,

the argument is that NSSE serves as an accurate measure of first-generation college students and data collected using this instrument can be used to make decisions to improve the college experience and therefore the success of first-generation college students. In framing evidence for validity this way, the argument for validity is no longer a Sisyphean task extending endlessly, but rather a constrained argument in a given context with clear parameters, which can be used as a building block with other arguments for validity over time (Kane, 2013).

National Survey of Student Engagement

The National Survey of Student Engagement is one of the most widely used and researched surveys in higher education, reaching 1,600 institutions and six million students since 2000 (Campbell & Cabrera, 2011; LaNasa et al., 2009; Kuh, 2008, Kuh, 2009; Porter, 2011; National Survey of Student Engagement, 2019; Ouimet, et al., 2004; Porter, 2011). The purpose of the survey is to provide high-quality, actionable data for student learning outcomes to inform institutional improvement, to discover and document effective practices in higher education, and to advocate for empirically derived indicators of collegiate quality (p. 10). One of the reasons for NSSE's popularity is because it emerged as an avenue to measure quality independent of "ratings, rankings, and reputations" (Thelin, 2019, p.147) and focused on aspects of the student experience, and the quality of these experiences, that institutions could concretely influence (Campbell & Cabrera, 2011; Kuh, 2009). Campbell and Cabrera (2011) explained that NSSE "had the potential to advance our understanding of the role of various student experiences (e.g., experiences with faculty, rigorous coursework, involvement in student organizations) in

collegiate outcomes (such as persistence and learning)" (p. 78). Measures of student engagement related to persistence and retention, as well as measures of educational and institution quality, also served as evidence to respond to accountability demands (Campbell & Cabrera, 2011).

Grounded in decades of research on student involvement (Astin, 1984) and integration (Tinto, 1975) NSSE focuses on student engagement in educationally purposeful activities and the educational effectiveness of the institution (NSSE, 2018). Engagement is not designed as a single construct, but a number of ideas grounded in research that relate to how college influences student learning and development (NSSE, 2018). Engagement reflects the choices and effort of students to engage with the institution as well as faculty and institutional efforts to actively engage students. In other words, engagement is conceptualized as a two-way street in which both the students and the institution have a responsibility for student success. In the conceptual framework for NSSE, Kuh wrote, "student engagement integrates what has been learned about quality of student effort, student involvement, and principles for good practice in undergraduate education into a broad framework for assessing quality and guiding its improvement" (NSSE, 2018, p. 42). The items and underlying constructs measured on NSSE were built around this idea of engagement. As such, data from NSSE reflect the commitment of students and the institution to student success.

During the design and testing of NSSE, Ouimet, Bunnage, Carini, Kuh, and Kennedy (2004) conducted focus groups, used expert advice, and cognitive interviewing to establish validity. In their study, Ouimet et al. (2004) hosted between three and six

focus groups at eight different institutions that had administered the NSSE. A total of 221 students participated in the 35 sessions of focus groups. In their reporting of the demographics of students participating in the focus groups, the authors cited the percent which were female or male, the class year of the participants, and noted that 37% were people of color (Ouimet et al., 2004, p. 237). In addition, researchers facilitated cognitive interviews with 28 undergraduates split equally between men and women. When discussing the findings and the changes made to the survey to increase validity, the authors noted that the findings from the cognitive interview showed "the majority of students interpreted the questions in identical or nearly identical ways" (Ouimet et al., 2004, p. 247). They also concluded that even if surveys have been explored with respondents to ensure proper interpretation, the survey should still be critically examined with diverse populations (Ouimet et al., 2004, p. 247).

The construct validity of NSSE was tested using two separate samples, one for exploratory factor analysis and one for confirmatory factor analysis. Confirmatory factor analysis was tested using the ten engagement indicators organized into the four themes and evaluated using multiple different fit indices (Miller et al., 2016). The fit criteria used were the chi-square error of approximation, goodness of fit index, comparative fit index, and the root mean square error of approximation (Miller et al., 2016). Overall fit indices, factor correlations, and regression weights provided evidence of construct validity for first-year students in reflective and integrative learning, higher-order learning, quantitative reasoning, learning strategies, collaborative learning, discussions with

diverse others, student-faculty interactions, effective teaching practices, quality of interactions and supportive environment (Miller et al., 2016).

In response to criticisms and questions of validity (Campbell & Cabrera, 2011; LaNasa et al., 2009; Porter, 2011), NSSE released an updated version in 2013 with an accompanying psychometric profile offering evidence of reliability and validity. To develop this updated version, NSSE staff engaged with campus users, reviewed the literature, examined existing data, hosted focus groups and cognitive interviews, and conducted psychometric testing (NSSE, 2018). The 2013 NSSE survey moves away from measuring five benchmarks towards ten engagement indicators of higher-order learning, reflective and integrative learning, learning strategies, quantitative reasoning, collaborative learning, discussions with diverse others, student-faculty interactions, effective teaching practices, quality of interactions and supportive environment (NSSE, 2018).

In the psychometric profile established by NSSE, Miller et al. (2016) conducted a confirmatory factor analysis across four groups of students and found evidence of model fit for each of the ten scales and their four related themes. Research conducted by Zilvinskis et al. (2017) supported the convergent and discriminant validity of the revised survey. Outside of the psychometric portfolio, few studies have examined the structure of NSSE, or the performance of NSSE across diverse groups of students.

Data from NSSE have been used to compare and benchmark student responses across institutions and to analyze subgroup data within institutions to identify areas of improvement (Fosnacht & Gonyea, 2018; NSSE, 2018). NSSE data has also been used as

a proxy for student learning outcomes such as critical thinking, communication, understanding society and culture, and civic engagement (Kuh, 2001). In a 2007 research study using a past version of NSSE data, focus groups and cognitive interviews, researchers found that educationally effective practices fostered desirable outcomes for all students, and historically underserved students (Black and Hispanic students in this study) benefited more than their White counterparts by engaging in such activities (Kuh et al., 2007). The same study reports that NSSE works equally well across students with differing racial and ethnic backgrounds (Kuh et al., 2007). This study did not include an examination of first-generation college students.

A gap in the research is the cultural responsiveness and invariance of the NSSE for first-generation college students. Given that NSSE has been used to make claims about the experiences of underrepresented students and how institutions can work to support the success of all students, the lack of interrogation of NSSE as a tool which is culturally responsive and equivalent for first-generation college students is significant. Researchers examining past and present versions of NSSE advocate for the continued collection of validity evidence across diverse populations (Fosnacht & Gonyea, 2018; Kuh, 2009; Ouimet et al., 2004). Additionally, NSSE is a tool designed and maintained largely by Dr. Kuh, a White male, built on theory developed by White men (e.g., Astin, 1984; Tinto, 1975). Using NSSE data as evidence of the quality of higher education and the basis for improving student outcomes under the assumption that the tool is responsive and invariant for historically marginalized students is a concrete example of centering privileged identities and ways of knowing. Taking a culturally responsive approach, this

study examines this assumption by questioning if a tool built and normed on predominantly White populations is culturally responsive and invariant, and therefore can be used to improve outcomes for historically marginalized students.

First-generation College Students

Establishing first-generation college students as a cultural population in higher education requires an overview of their demographic characteristics, educational outcomes, and shared lived experiences in higher education. First-generation college students are racially and socio-economically diverse students, historically underrepresented and underserved in higher education. Means and Pyne (2017) write, "low-income, first-generation college students and Students of Color bring identities that have been historically outside or invisible within higher education" (p. 921). In 2012, 24 percent of college students were first-generation (Redford & Hoyer, 2017). Of these students, 27 percent identified as Hispanic or Latino, 14 percent as Black or African American, and five percent as Asian (Redford & Hoyer, 2017). Half of first-generation college students reported their household income at between $20,001 and $50,000, and 27 percent reported a household income of $20,000 or less (Redford & Hoyer, 2017).

Prior research has shown that first-generation college students experience differential outcomes and experience the higher education environment differently than their continuing-generation peers (Cataldi et al., 2018; Gibbons et al., 2019; Pike & Kuh, 2005; Redford & Hoyer, 2017). First-generation students experience differential outcomes compared to their continuing-generation peers in terms of persistence and graduation (Catalidi et al., 2018; Pascarella et al., 2004; Radunzel, 2018). Persistence

rates for first-generation and continuing-generation college students are showcased in Table 3. After enrolling in higher education, one-third of first-generation students left without earning a degree compared to 14 percent of continuing-generation students (Cataldi et al., 2018, p. 9). After six years, 56 percent of first-generation students earned a credential or degree compared to 74 percent of continuing-generation students (Cataldi et al., 2018, p. 9). Differences in outcomes extend to engagement and sense of belonging (Pike & Kuh, 2005). Pike and Kuh (2005) used multi-group structural equation models to examine if first-generation college student experiences affected their learning and development in higher education compared to their continuing-generation peers. They found that first-generation and continuing-generation college students varied significantly in their college experiences and learning outcomes including academic and social engagement, and sense of integration or belonging.

Table 3: Academic Progress (Cataldi et al., 2018)

	First-generation College Students	**Continuing-generation College Students**
Leave higher education without a degree	33%	14%
Earn a degree or credential after six years	56%	74%

Many first-generation college students have lived experiences which pose challenges to their success such as less implicit knowledge of higher education environments, navigating power dynamics of faculty, the cost of higher education, and ongoing commitments to family. Gibbons, Rhinehart, and Hardin (2019) conducted focus groups with first-generation college students and found that barriers to students' success

include family, finances, and a lack of information about how to navigate higher education (e.g. financial aid, purposeful educational activities).

Pike and Kuh (2005) reported that one reason first-generation students may have different engagement experiences is that they may not understand the importance of campus engagement, they write, "compared to second-generation college students, they have less tacit knowledge of and fewer experiences with college campuses and related activities, behaviors, and role models" (p. 290). Like Pike and Kuh (2005), Pascarella et al. (2004) found:

> Those with college-educated parents have better access to human and cultural capital through family relationships. Consequently, compared to their peers with highly educated parents, first-generation students are more likely to be handicapped in accessing and understanding information and attitudes relevant to making beneficial decisions about such things as the importance of completing a college degree, which college to attend, and what kinds of academic and social choices to make while in attendance. (p. 252)

Cultural capital includes information required to be successful in higher education passed down from parents to their children over time (Ward et al., 2012). Without this implicit knowledge, studies have found that first-generation college students are unaware of the impact of educationally purposeful activities and, as such, are less likely to be engaged in on campus activities, live on campus, have positive perceptions of the campus climate, or perceive faculty as caring about their success (Means & Pyne, 2017; Pascarella et al., 2004; Pike and Kuh, 2005).

Interactions with faculty also influence first-generation college student experiences. Means and Pyne (2017) found that "faculty decisions about pedagogical

approach, attitude towards students in class and during office hours, and expectations of student knowledge, behavior and ability frequently shaped student choices about how to respond to a class, a task, or a major" (p. 917). Some first-generation students found it easy to approach faculty while others found approaching faculty, asking questions, and attending office hours to be an intimidating experience (Means & Pyne, 2017). In addition, many first-generation students reported, "being uncomfortable with struggling when they perceived their peers as successful" (Means & Pyne, 2017, p. 917). This combination of faculty and peer experiences can leave first-generation college students feeling isolated and "behind" in navigating the college environment.

First-generation college students are more likely than their continuing-generation peers to leave higher education due to costs (Redford & Hoyer, 2017). To pay for higher education, first-generation students often work (Redford & Hoyer, 2017), which impacts their ability to engage in educationally purposeful activities as measured by NSSE. Students who are the first in their families to go to college also leave higher education because of changes in family status, conflicts with demands at home, and personal problems at higher rates than their continuing-generation peers (Redford & Hoyer, 2017). First-generation college students no longer fit neatly into the world they left, with immediate family members who did not attend higher education, and as a result they struggle to articulate challenges with their transition to their family members (Gibbons et al., 2019).

Although prior studies take a deficit orientation to describing first-generation college student experiences (e.g. Pascarella et al., 2004), students who are the first in

their family to attend college find strength in their family support systems, are resilient and growth oriented. Families are also a strong area of emotional support and encouragement for students (Gibbons et al., 2019). Gibbons et al. (2019) found that even if families did not fully understand the challenges faced by their students, they still provided a strong foundation of emotional support and encouragement for success. First-generation students are resilient and growth oriented. Gibbens et al. (2019) found that students actively prepared themselves for change in adjusting to a new environment, preparing for academic rigor, and balancing multiple responsibilities.

While research has shown that first-generation students experience higher education differently and face a number of challenges compared to their continuing-generation peers, studies also show that successful persistence and graduation can have an equalizing effect on educational and employment outcomes. One study found that when first-generation college students do engage in purposeful activities, they, "tended to derive significantly stronger positive benefits from these involvements than did other students" (Pascarella et al., 2004, p. 273). These findings suggest that engaging first-generation college students in the higher education environment is especially important for their success. Cataldi et al. (2018) found no significant differences between first-generation and continuing-generation college student full-time employment levels four years after graduation. They also found no differences in median salaries between first and continuing-generation graduates (Catalidi et al., 2018). After graduation, both groups were equally as likely to pursue graduate and doctoral education (Cataldi et al., 2018). First-generation students are experiencing and engaging in the university context in

different ways than their continuing-generation peers, but when they persist and graduate, they experience the same post-graduate outcomes as their continuing-generation peers.

Given the demographic characteristics and lived experiences of first-generation college students, there are reasons to believe NSSE may not be culturally responsive nor invariant. NSSE is administered in students' first and fourth years in the higher education environment. For first-generation college students, the first year is one of significant challenge and adjustment (Ward et al., 2012). By virtue of being the first in their family to attend higher education, first-generation college students do not have the same cultural capital as their continuing-generation peers (Pascarella et al., 2004; Ward et al., 2012) and may not recognize the activities examined on NSSE as educationally purposeful. Examples of such items include asking peers for help or attending faculty office hours. Research has shown first-generation college students as hesitant to do either (Means & Pyne, 2017; Pascarella et al., 2004; Pike & Kuh, 2005). Studies have used NSSE data to draw conclusions and make decisions for historically marginalized students, and the survey was designed to provide information about education quality (NSSE, 2018), but the survey has not been validated for such use with first-generation students.

Culturally Responsive Evaluation as a Theoretical Lens for this Study

In the midst of a social awakening (mostly for White people) with police brutality leading to the deaths of hundreds of Black and Brown human beings, all recorded and viewed by a world-wide audience, people are protesting and fighting to challenge the foundation of American society, to unearth and expose oppressive ideologies that undergird every system and structure from police to education. As a White woman in

education drawn to culturally responsive work, choosing a framework for this research which fosters transformation, attends to oppressive histories, and centers cultural ways of knowing and being is an apt fit. Culturally responsive evaluation is a transformative approach which centers culture throughout the evaluation process, it is also a philosophy and guiding framework. Hopson (2009) argues, "an important theoretical and historical development of CRE is its stance to challenge and resist dominant, mainstream thinking that pervades the tradition of scholarship that sees difference as deficit or diversity as deviant" (Hopson, 2009, p. 441). The transformational evaluation process is an opportunity to push back on dominant discourse, decenter Whiteness, and reposition diverse lived experiences and knowledge as expertise in the collection of credible evidence in determining the merit or worth of a program.

Dahler-Larson's (2012) uses the metaphor of an "evaluation machine" to describe a dystopian evaluation process which blindly feeds the demand for data to inform policy. This notion of an evaluation machine thrives on efficiency and repetition, maintained by the professionals operating within the machine. Efficiency and repetition are robustly manifested in the routinization of methodological selection without critically examining whether such tools produce data which are accurate in a given context, and for given populations. One critical problem resulting from an overemphasis on efficiency and repetition is that this focus has the propensity to reduce complex social issues into performance indicators that hold little weight and are not designed by the people they are intended to measure. Dahler-Larson (2012) warns against such routinization by saying, "used prescriptively, simplistic models can have unpleasant and inappropriate

consequences in complex reality" (p. 37). An example of such prescriptive use of methods is readily available in higher education. The same measurements have been used for decades without critical re-evaluation to center marginalized perspectives (e.g., IPEDs reporting, National Survey for Student Engagement). Simultaneously, while the demographics of students have changed drastically, significant attainment gaps persist. The entrance and subsequent withdrawal of historically marginalized students without completing a degree is contributing to the national debt crisis related to education and presents, in the media and in national discourse, as an ineffectiveness of higher education. This study steps outside the routine use of established measures using a culturally responsive evaluation lens to attend to the gap between the measurement tools and the students they are designed to measure to position cultural responsiveness and measurement invariance as aspects of multicultural validity evidence for the use of NSSE with first-generation college students in the context of higher education.

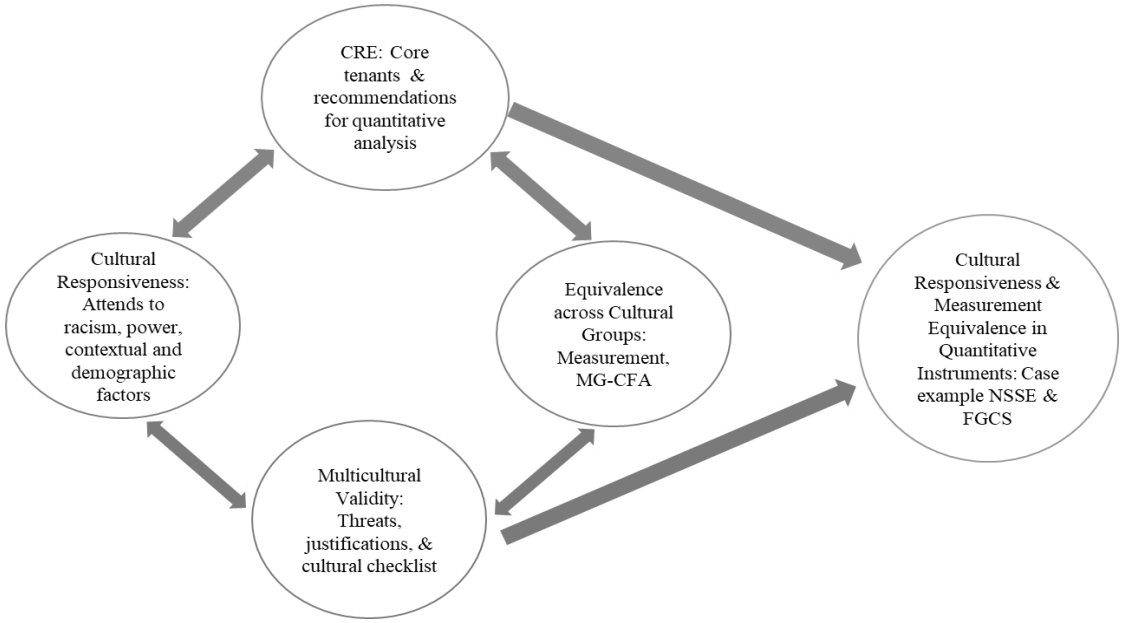
Five tenants are core to culturally responsive evaluation practice: evaluator positionality, the role of evaluators in furthering social change by challenging systems of inequality, embracing stakeholders and their cultural perspectives, centering culture throughout the evaluation process, and the uplifting contributions of culturally and ethnically diverse communities in instigating change (Hopson, 2003). Culturally responsive approaches address context, fully attend to diverse voices, power, identity demographics, socioeconomic status and the sociopolitical context (Stokes et al., 2011). Recommendations for the use of quantitative approaches to data collection and analysis stemming from the foundational literature on CRE include engaging in reflexivity and

representation of voice, establishing equivalence and validating instruments in specific contexts, considering contextual conditions in data analysis, conducting analysis between and within groups, and conducting multiple forms of analysis. Traditional statistical techniques require large sample sizes, focus on the averages, and facilitate comparisons across group performances. Given the lens of the researcher, these comparisons are often from underrepresented groups to the White majority population. The research process is shaped by the positionality of the researcher which influences data collection, analysis, and interpretation (Frierson et al., 2010; Hood et al., 2015; Hughes et al., 1993). Using statistical techniques requires decisions influenced by the researcher's positionality with implications for equity. Focusing on large samples sizes often automatically removes smaller populations in higher education, such as Native Americans, from the discussion all together or requires choices around aggregating smaller groups into larger groups (e.g. "students of color") to include diverse perspectives in analysis, but still effectively removes unique experiences across and within smaller groups.

This study positions cultural responsiveness and invariance of a standardized measure as evidence of multicultural validity. Kirkhart (2013) provides a checklist inclusive of history, location, power, voice, relationships, time, reciprocity, plasticity and reflexivity in the argument for multicultural validity. Kirkhart (2013) also outlines five areas of justification or threats to validity including methodological, experiential, relational, theoretical and consequential. Using techniques such as multi-group confirmatory factor analysis to examine invariance shifts the analysis from student performance, to test performance for students.

Figure *2* below show how the different aspects of culturally responsive evaluation, measurement, and multicultural validity interact to provide the theoretical lens for this study using the National Survey of Student Engagement and first-generation college students as a case study.

Figure 2: Culturally Responsive Evaluation as a Theoretical Lens



Summary of the Literature

Hood, Hopson, and Kirkhart (2015) question how evaluators know if they are culturally responsive and provide a framework for culturally responsive evaluation. One aspect of this framework emphasizes the importance of multicultural validity (Kirkhart, 1995) and focuses on the methodologies and methods used for data collection and how these yield data that are valid across marginalized populations. Prior research in CRE found that evaluators using culturally responsive approaches often used qualitative and

mixed methods and drew attention to the lack of quantitative instruments designed and normed with cultural groups (Chouinard & Cousins, 2009; Frierson et al., 2002). Dominant perspectives are reinforced or disrupted through methodological choices. The lack of available instruments normed for marginalized populations inhibits the ability of evaluators to draw valid inferences about the needs and experiences of marginalized populations nor to disrupt dominant perspectives.

In an effort to attend to historical context and center cultural perspectives, culturally responsive evaluators using quantitative measures must interrogate assumptions that such measures collect valid, accurate, and trustworthy evidence across cultural contexts and populations (AEA, 2011; Hood et al., 2015). Establishing invariance is a repeated concern noted for evaluators using a culturally responsive approach and is a form of analysis which establishes lack of bias in quantitative measures. Invariance across cultural groups serves as evidence of data validity and lack of bias in standardized measures. Using techniques such as multi-group confirmatory factor analysis shifts the analysis from student performance, to test or assessment performance. The *Standards for Educational and Psychological Testing* (1999) encourage expanding the examination of outcomes across subgroups beyond race and gender. This study responds to this call by focusing on first-generation college students using data from NSSE.

How students engage in the context around them and interact with power dynamics, systems, and structures engrained in higher education are molded by individual and systemic factors that are shaped by culture. Drawing on culturally

responsive evaluation as a lens, this study examines approaches to quantitative and mixed methods of data collection in both culturally responsive evaluation and measurement across cultures to determine how measures are identified as culturally responsive and invariant. The purpose of the current study is to combine strategies used in these two fields in a single case study using the National Survey of Student Engagement and first-generation college students. This research study questions the assumption that a measure normed on a predominately White, continuing-generation population is an accurate, trustworthy, and culturally responsive measure for historically marginalized students.

CHAPTER III

METHODOLOGY

In the current study, I examined approaches to quantitative and mixed methods data collection and analysis in the fields of culturally responsive evaluation and measurement in order to identify and employ strategies to determine to what extent the National Survey of Student Engagement is a culturally responsive and invariant measure for first-generation college students. To this end, I asked the following research questions:

1. To what extent is the National Survey of Student Engagement a culturally responsive measure for first-generation college students? What considerations for rendering quantitative measures culturally responsive can be derived from a critical examination of the empirical literature on culturally responsive evaluation and measurement?
2. To what extent do statistical techniques used in measurement, such as multi-group confirmatory factor analysis, establish measurement invariance in the National Survey for Student Engagement for first-generation college students? How does this approach further cross-cultural understanding?

Research Design

The central problem addressed in this study was identifying how quantitative surveys could be evaluated as culturally responsive and invariant, centering first-

generation college students in an examination of the National Survey of Student Engagement as a case example. Mixed methods are often used in a research design when quantitative or qualitative methods independently would not yield sufficient evidence to answer the research questions (Creswell, 2014). The research design for this study was a convergent parallel mixed-methods design in which two questions were answered independently, using separate methodologies, and findings from each study converged to provide a comprehensive analysis with which to respond to the research problem. Table 4 below outlines the research questions, the methods used to answer these questions, and associated data sources.

Table 4: Summary of Research Questions, Methods, and Data Sources

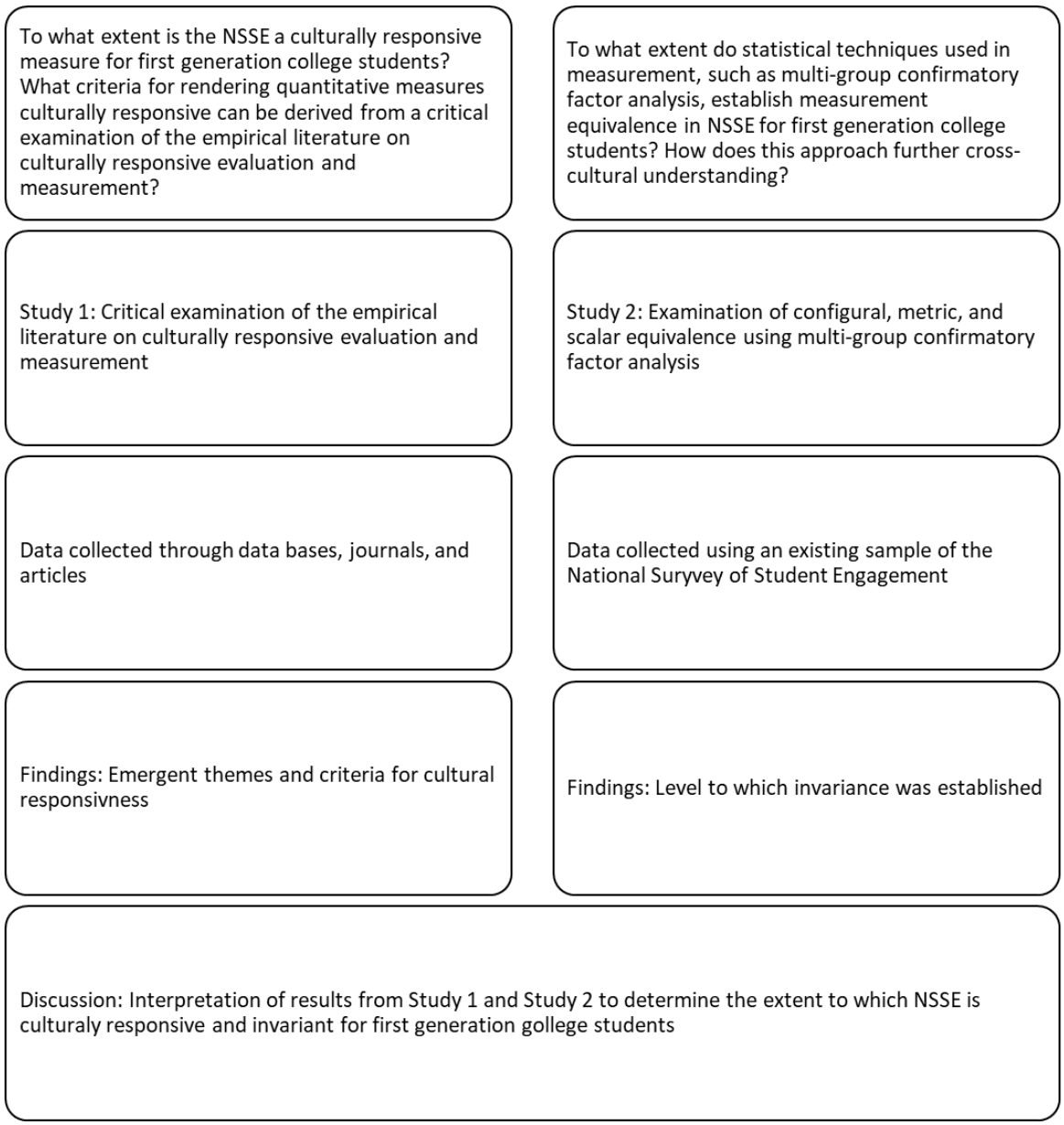
Research Question	**Method**	**Data Source**
1. To what extent is the National Survey of Student Engagement a culturally responsive measure for first-generation college students? What considerations for rendering quantitative measures culturally responsive can be derived from a critical examination of the empirical literature on culturally responsive evaluation and measurement?	Critical literature review	Journal articles
2. To what extent do statistical techniques used in measurement, such as multi-group confirmatory factor analysis, establish measurement invariance in the National Survey for Student Engagement for first-generation college students? How does this approach further cross-cultural understanding?	Multi-group confirmatory factor analysis (MG-CFA)	National Survey of Student Engagement

The combination of a critical literature review and statistical analysis is a form of mixed-methods research. Greene (2007) approaches mixed methods as a way to actively and intentionally engage in difference by respecting multiple ways of knowing and engaging in methodological diversity. She writes that including multiple methodologies and methods, "…enhances not only the generative potential of mixed methods inquiry but also its potential to respect, appreciate, and accept variation and diversity in the substance of what is being studied" (p. 28). Furthering this idea, Mertens (2007, 2011) positions the use of culturally competent mixed methods in the transformative paradigm as advancing social justice and social change. In explaining the value of mixed-methods in the social justice branch of evaluation, Mertens (2018) writes, "the use of mixed methods designs in evaluation rooted in the Social Justice branch allows for the capture of different realities in their complexity from the view of stakeholders' lived experiences" (p. 21). This study proposed the intentional integration of both quantitative and qualitative methods of data collection to yield insights regarding to what extent NSSE is a culturally responsive and invariant measure for first-generation college students.

The convergent parallel design included a critical literature review and a multi-group confirmatory factor analysis. *Study one* was a critical literature review focused on identifying how empirical articles in culturally responsive evaluation and measurement establish quantitative measures as culturally responsive and invariant. Themes as well as gaps were examined. Findings from this study were used to develop core considerations for examining if a quantitative measure is culturally responsive. I used the critical considerations established from study one to inform an evaluation of whether NSSE was

culturally responsive. *Study two* involved a multi-group confirmatory factor analysis to determine the extent to which configural, metric, and scalar invariance across first-generation and continuing-generation students could be established. Criteria developed from study one and invariance results from study two converged in the discussion to determine to what extent NSSE was both culturally responsive and invariant for first-generation college students. An illustration of this mixed-methods design consisting of two studies is illustrated in Figure 3.

Figure 3: Research Design



Study One: Critical Literature Review

Study one was a critical examination of the empirical literature in culturally responsive evaluation and measurement. The literature was examined to identify strategies (e.g., stakeholder engagement, advisory boards or review panels, multi-group confirmatory factor analysis) for how these two fields establish cultural responsiveness and invariance across diverse groups. Findings from the critical review were used to evaluate NSSE as a culturally responsive survey. The purpose of the comprehensive review of the literature was to examine the landscape of how quantitative surveys are determined to be culturally responsive and invariant when used with cultural populations. How do culturally responsive evaluators navigate the use of standardized measure with culturally diverse populations? What approaches do culturally responsive evaluators take to using quantitative surveys in cultural populations? How do they determine if surveys are culturally responsive? Do they attend to equivalence/invariance? In the measurement literature, how was culture addressed or included in the analysis of equivalence/invariance?

Sample Selection and Data Collection

I examined empirical articles from 2000-2019 for study one. Articles included evaluation and measurement studies with quantitative or mixed-method approaches, as well as reflections on past work, or case studies on prior work. Search terms for the critical literature review include “culture,” “cultural,” “cross-cultural,” "culturally responsive," “bias,” "fairness," and “equivalence.” Inclusion criteria for articles considered in study one included detailed attention to cultural identities or cultural

groups, two or more of the search terms listed above, detailed description of data collection or analysis with diverse populations, and discussions of cultural responsiveness or equivalence. The nuance of each of these individual search terms as well as their various combinations is significant as a search for "bias" in a measurement journal yielded an unwieldy amount of results many of which were not directly connected with the research questions posed in this study. The listed terms were used to search *American Journal of Evaluation, Evaluation, American Journal of Community Psychology, Canadian Journal of Program Evaluation, Evaluation and the Health Profession, Journal of Multidisciplinary Evaluation, New Directions for Evaluation, and Studies in Educational Evaluation*. These terms were also used to search *Psychometrika, Journal of Educational Measurement, Educational Measurement: Issues & Practice, Educational and Psychological Measurement, Applied Psychological Measurement, Measurement and Evaluation in Counseling and Development, Journal of Educational and Behavioral Statistics, International Journal of Testing,* and the *Journal of Method and Measurement in Social Sciences*. In addition, bibliographies, books, and book chapters were examined for additional articles.

Review Strategy and Analysis

Articles which met the inclusion criteria were read repeatedly, analyzed and summarized in two ways. First, a one-page summary of notes was created for each article to allow for an in-depth, rich qualitative analysis. Second, each article was cataloged as a row of data in a spreadsheet. Included in the summary table were descriptions of the program, descriptions of the study sample methods of data collection, analysis strategies,

and discussions of responsiveness and/or equivalence. Once the summary table was completed, findings were synthesized and analyzed for themes and patterns across studies. The themes frame an understanding of the strategies used across the two fields and identify critical considerations for imagining quantitative surveys as culturally responsive.

Qualitative research is emergent, reflexive, and makes use of inductive and deductive data analysis techniques (Creswell, 2014). As such, a grounded theory approach to qualitative research guided the data analysis. Grounded theory allows for the identification of general themes and can offer new insights in the study of a specified phenomenon, specifically, this approach allowed for meaning to be drawn from the data, not prior to the data collection (Corbin & Strauss, 2014). Data from the summaries and matrix were analyzed using a constant comparative approach (Corbin & Strauss, 2014). Creswell (2014) describes inductive and deductive analysis: "qualitative researchers build their patterns, categories, and themes from the bottom up by organizing the data into increasingly more abstract units of information" (p. 186). A constant comparative approach allows for this inductive and deductive process to unfold during data analysis. Analyzing qualitative data is an iterative process, shaped by researcher positionality and reflexivity. The first part of my iterative process was reviewing the articles and writing a synopsis of each article. In a constant comparative approach, data are analyzed through constant comparisons where one small piece of data is compared for similarities and differences, data similar to one another are grouped together under a theme, and the emergence of multiple themes provides a foundation to for drawing conclusions (Corbin

& Strauss, 2014). Notes on each article were analyzed and compared to one another to identify emerging themes and draw conclusions across the two bodies of literature.

Study Two: Multi-Group Confirmatory Factor Analysis

Study two was a causal-comparative study examining to what extent the National Survey of Student Engagement was invariant for first-generation and continuing-generation college students. Causal-comparative studies are comparisons of groups based on characteristics that cannot be altered such as race, or first-generation college student status (Mertens, 2009). This study used multi-group confirmatory factor analysis as a technique to compare to what degree NSSE was invariant for first-generation college students and continuing-generation college students. Three levels of invariance were examined configural, metric, and scalar.

Sample Selection

Data for this study were drawn from an existing sample of NSSE survey respondents. The criteria for inclusion were first-year, first-generation college students, between the ages of 18-22, who participated in NSSE 2016 or 2017. Additionally, a sample of first-year, continuing-generation college students between the ages of 18-22 was also drawn. The sample was randomly selected from respondents participating in NSSE 2016 or 2017, attending doctoral-granting research institutions with enrollment of 20,000 students or more. Students who may have attended another institution and transferred, were not included in the sample for this study. On the NSSE questionnaire designed for first-year students, the following question was asked: “What is the highest level of education completed by either of your parents (or those who raised you)?” with

response options including “Did not finish high school” and “High school diploma or G.E.D”. These options are congruent with how this study defined first-generation college students. In total, 3,000 first-generation college students and 3,000 continuing-generation college students were included in the study.

Instrumentation: National Survey of Student Engagement

The National Survey of Student Engagement has been developed and tested for over twenty years but has not been examined for cultural responsiveness or bias for first-generation college students (FGCS), a population historically under-served in higher education. At the time of development and piloting the survey, first-generation college students were not explicitly considered and searches for current studies specifically focused on NSSE and first-generation college students yield few relevant results. Studies which have examined the qualities of NSSE have concluded with the need to examine the survey in more detail across cultural groups (e.g., Kuh, 2009; Ouimet et al., 2004). The NSSE measures ten engagement indicators (EIs) with 47 items. The factors/engagement indicators, items, and response options are outlined in Table 5.

Table 5: NSSE Themes, Factors, and Items

Engagement Indicator	Items	Response Choices
Higher-Order Learning	Apply facts, theories, or methods to practical problems or new situations Analyzing an idea, experience, or line of reasoning in depth by examining its parts Evaluating a point of view, decision, or information source Forming a new idea or understanding from various pieces of information	Very much, Quite a bit, Some, Very little

Reflective & Integrative Learning	Combined ideas from different courses when completing assignments Connected your learning to societal problems or issues Included diverse perspectives in course discussions or assignments Examined the strengths and weaknesses of your own views on a topic or issue Tried to better understand someone else's views by imagining…his or her perspective Learned something that changed the way you understand an issue or concept Connected ideas from your courses to your prior experiences and knowledge	Very much, Quite a bit, Some, Very little
Learning Strategies	Identified key information from reading assignments Reviewed your notes after class Summarized what you learned in class or from course materials	Very often, Often, Sometimes, Never
Quantitative Reasoning	Reached conclusions based on your own analysis of numerical information Used numerical information to examine a real-world problem or issue Evaluated what others have concluded from numerical information	Very often, Often, Sometimes, Never
Collaborative Learning	Asked another student to help you understand course material Explained course material to one or more students Prepared for exams by discussing or working through course material w/other students Worked with other students on course projects or assignments	Very often, Often, Sometimes, Never
Discussions with Diverse Others	Discussions with… People of a race or ethnicity other than your own Discussions with… People from an economic background other than your own Discussions with… People with religious beliefs other than your own Discussions with… People with political views other than your own	Very often, Often, Sometimes, Never
Student-Faculty Interaction	Talked about career plans with a faculty member Worked with a faculty member on activities other than coursework Discussed course topics, ideas, or concepts with a faculty member outside of class	Very often, Often, Sometimes, Never

	Discussed your academic performance with a faculty member	
Effective Teaching Practices	Instructors… Clearly explained course goals and requirements Instructors… Taught course sessions in an organized way Instructors… Used examples or illustrations to explain difficult points Instructors… Provided feedback on a draft or work in progress Instructors… Provided prompt and detailed feedback on tests or completed assignments	Very much, Quite a bit, Some, Very little
Quality of Interactions	Quality of interactions with… Students Quality of interactions with… Academic advisors Quality of interactions with… Faculty Quality of interactions with… Student services staff… Quality of interactions with… Other administrative staff and offices…	1 = Poor to 7 = Excellent, Not Applicable
Supportive Environment	Inst. emphasizes… Providing support to help students succeed academically Inst. emphasizes… Using learning support services Inst. emphasizes… Encouraging contact among students from different backgrounds Inst. emphasizes… Providing opportunities to be involved socially Inst. emphasizes… Providing support for your overall well-being… Inst. emphasizes… Helping you manage your non-academic responsibilities Inst. emphasizes… Attending campus activities and events Inst. emphasizes… Attending events that address important social/econ./polit. issues	Very much, Quite a bit, Some, Very little

The reliability of each of the scales on NSSE is provided in Table 6. Using the Cronbach's alpha as an indicator of internal consistency, alpha was over .80 for each of the scales with the exception of the learning strategies scale (NSSE, 2019).

Table 6: Reliability of NSSE Indicators (NSSE, 2019)

Engagement Indicator	Cronbach's Alpha: First-year students
Higher-Order Learning	.83
Reflective & Integrative Learning	.85
Learning Strategies	.76
Quantitative Reasoning	.82
Collaborative Learning	.83
Discussions with Diverse Others	.87
Student-faculty Interaction	.81
Effective Teaching Practices	.84
Quality of Interactions	.85
Supportive Environment	.88

Data Collection

Data for the quantitative portion of this study had already been collected by colleges and universities across the United States and stored with the National Survey of Student Engagement Institute. To obtain a sample of data from the NSSE institute, a request for research form was submitted, approval granted, and the specified sample of data distributed to the researcher.

Data Analysis

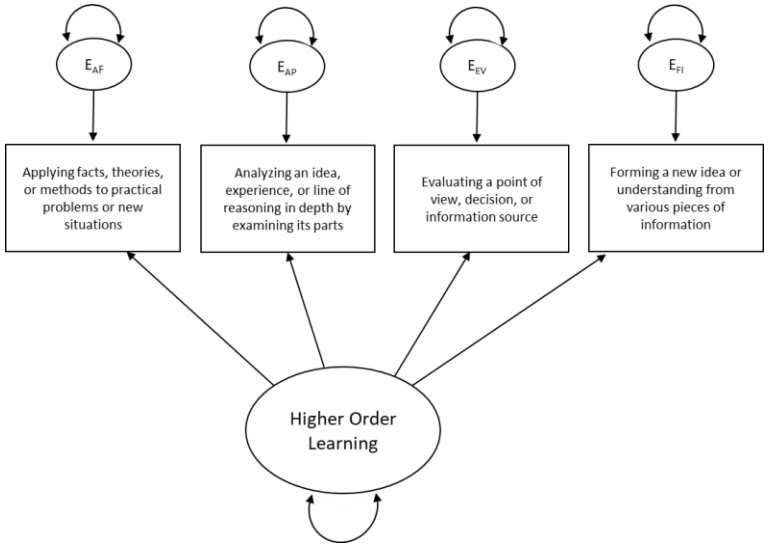
I used structural equation modeling techniques to determine to what extent NSSE held configural, metric, and scalar invariance for first-generation and continuing-generation college students. Structural equation modeling is a family of statistical techniques that can be used to specifically quantify and test a theory of interest using both observed (indicators) and latent variables (factors) (Kline, 2011). One of the benefits of SEM techniques is that they account for residual (error) terms associated with observed or latent variables, and in the case of observed variables, this residual term represents the

unexplained variance in an indicator and its related factor. Confirmatory factor analysis is a measurement model in structural equation modeling and a multi-sample, or multi-group, confirmatory factor analysis model is a measurement model fitted to data for more than one group at the same time. Results from such a model provide evidence of measurement invariance (Kline, 2011). No studies have applied multi-group confirmatory factor analysis to the study of first-generation college students using data from NSSE. The model proposed in this study builds off of past research conducted by Miller et al. (2016) regarding the factor structure of NSSE and expanded their research by confirming the factor structure across a sample of first-generation college students and continuing-generation college students and examining to what extent configural, metric, and scalar invariance could be established for both groups.

Testing for invariance is a nested process in which factor loadings, thresholds, and error variances were constrained or freely estimated. To test configural invariance, a model built off of the first order models specified by Miller et al. (2016) was specified and fit simultaneously to the sample of first-generation college students and continuing-generation college students, keeping the number of factors and their associated indicators the same, while freely estimating all parameters (Byrne et al., 1989; Cheung & Rensvold, 2002). In establishing configural invariance, the same pattern of loadings should occur, but the weight of these loadings can differ, so the researcher can conclude that the same constructs are present for both groups (Kline, 2011). Figure 4 provides an illustration of the expected structure for the factor of higher order learning and associated items. This

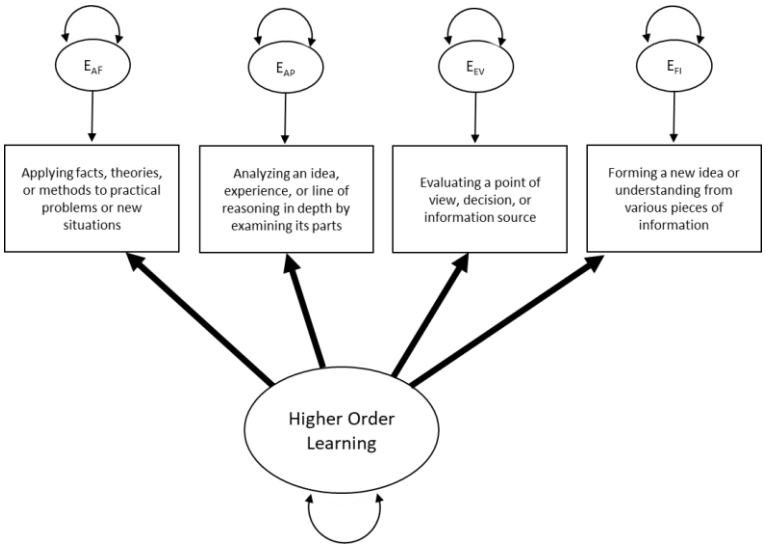
same factor structure should hold across both student groups. The strength of the relationships in this model were freely estimated as were the thresholds.

Figure 4: Configural Invariance



To test metric invariance, factor loadings were constrained for each indicator across groups while the intercepts/thresholds were freely estimated (Byrne et al., 1989; Cheung & Rensvold, 2002). In this case, the factor loadings were fixed to be the same, but the intercepts/thresholds were allowed to differ and one group was able to have more or less measurement error than the other. Intercepts were allowed to differ, affectively allowing for groups to differ in how difficult it was to endorse a specific item. Results from this level of analysis were compared to the configural analysis using a chi-squared test of difference and changes in CFI as indicators of model fit. Figure 5 provides an illustration of the metric model where the relationships between the items and the factor were fixed across both groups, as indicted by the bolded lines between factors and items.

Figure 5: Metric Invariance



Scalar invariance tests to see if the groups in the MG-CFA use the scale in the same way (Cambell et al., 2008). In the scalar model both factor loadings and intercepts were fixed. A conceptual illustration of scalar invariance is not pictured because showing constrained intercepts on a path diagram appears cumbersome and may add confusion rather than clarity. The test for scalar invariance determines if the test performs the same across groups and allows for the comparison of latent variable means. Results from the metric level of analysis were compared to the scalar level of analysis to determine changes in fit.

Several fit statistics were used to evaluate models fit in the multi-group confirmatory factor analysis. The first fit statistic was the model chi-square statistic, which is statistically significant at $p < .05$. The Root Mean Squared Error of Approximation (RMSEA) is a second fit statistic. General criteria for RMSEA are < .10 indicates marginal fit, < .08 indicates good fit, and < .05 indicates excellent fit. The third

and fourth indices are the Comparative Fit Index (CFI) and the Goodness of Fit Index (GFI), which both range from zero to one with one being the best fit. For both CFI and GFI, indices values greater than .90 indicate marginal to good fit and values greater than .95 indicate excellent fit. The Standardized Root Mean Squared Residual (SRMR) is the last fit index used, with values less than .08 generally considered to be an indication of good fit (Kline, 2011).

Data Quality

Validity, reliability, and trustworthiness of the data are markers of data quality. Data quality for quantitative data were established through the use of confirmatory factor analysis, Cronbach's alpha, and using multiple fit indices in the invariance testing. For qualitative data, thick descriptions and an audit trail were used to determine data quality. Each study was conducted separately and the findings from each study were triangulated and converged to allow a determination about the cultural responsiveness and invariance of the National Survey of Student Engagement for first-generation college students. Triangulating the findings between study one and study two as served as an additional opportunity to check data quality.

Quantitative Data

Quantitative data quality is concerned with internal validity and reliability of the data. Prior to conducting analysis, SPSS was used to screen data including checks for issues of multivariate normality, missing data, and multicollinearity. As a measure of internal consistency, Cronbach's alpha coefficient was calculated for all ten scales. The scales analyzed included higher order learning (HO), reflective and integrative learning

(RI), learning strategies (LS), quantitative reasoning (QR), collaborative learning (CL), discussions with diverse others (DDO), student-faculty interaction (SF), effective teaching practices (ET), quality of interactions (QI), and supportive environment (SE). As discussed in the analysis section, multiple fit indices were used to establish model fit. Past research to develop evidence of construct validity conducted on the newest form of NSSE (Miller et al., 2016) served as the foundation for building the confirmatory factor analysis for this study.

Qualitative Data

Congruent with a grounded theory approach and the constant comparative method, I used an audit trail and thick descriptions to establish credibility, transferability, dependability, and confirmability in the qualitative analysis (Lincoln & Guba, 1985). The combination of one page notes and the data matrix of synthesized findings allowed for the development of thick and rich descriptions providing an audit trail which, in the write up of the data, could help the reader to reach their own conclusions about the credibility and trustworthiness of the study (Maxwell, 2013). Thick descriptions include a "presentation of procedures, context, and participants in sufficient detail to permit judgement by others of the similarity to potential application sites; specify minimum elements necessary to 're-create' findings" (Nastasi & Hitchcock, 2016, p. 71).

Data Quality in Mixed Methods

In addition to establishing judgments of quality for both quantitative and qualitative data, specific points of data integration, data triangulation across sources and methods, served as another opportunity to establish data quality. Data triangulation is

defined as “data collection, analysis, and interpretation based on multiple sources, methods, investigators and theories” (Nastasi & Hitchcock, 2016, p. 70). Triangulation happened in this study as the data from the critical examination of the literature was used to evaluate NSSE as a culturally responsive tool, while the statistical analysis identified any areas in which the survey functioned equivalently across populations, and findings from the two studies converged to provide evidence and make a determination of the extent to which NSSE served as a culturally responsive and invariant measure for first-generation college students.

Researcher Positionality

The research problems examined in this study are motivated in part by many of the identities and experiences of myself as a researcher. As a young, White, woman growing up in rural Pennsylvania, I was awarded a full-ride scholarship through a private foundation to attend a small, private liberal arts college. My mother was a single parent who made $6.28 an hour while I was growing up. I was the first in my family, a first-generation college student, to go to school. I went on to obtain two master’s degrees, and pursue a doctorate in educational research. In 2002, the idea of first-generation college students was not a mainstream research topic emergent in higher education. Although I came from a different socio-economic status than the vast majority of students at the institution, my success was assumed. I had no idea of the importance of faculty interactions, working with students who were different from myself, or what higher order learning activities were. As an undergraduate, I had already developed a deep-rooted dedication to understanding diverse experiences, and as an international relations major

learned the systems and structures which uphold power and oppression and sustain injustice. With my Master's degree in higher education, I went on to work with students from diverse backgrounds and provide programs to support their success, all while operating within a context initially designed to serve an elite few.

These lived experiences, combined with education and professional experiences have converged in this doctoral work. Over the course of my studies, I have strengthened my assessment work with quantitative and qualitative analysis approaches and learned more about evaluation as a mechanism for change. As a doctoral student with 10 years of full-time experience working in higher education, my experiences have taught me about the relationships between culture, data, equity, and student outcome development. I have given conference presentations, written articles, sat on panels, offered trainings, and penned blog posts about the importance of cultural responsiveness in higher education assessment practices. Taking an approach to research which centers historically oppressed voices using quantitative instruments that are often upheld as the gold standard for data generalization during a time of social awakening and a global pandemic feels like small but meaningful thread to pull in becoming a researcher and evaluator who conducts work to make the world better for women and the global majority.

CHAPTER IV

RESULTS: STUDY ONE

The purpose of study one was to conduct a critical examination of the empirical literature in culturally responsive evaluation and measurement to explore how quantitative measures are determined to be culturally responsive and invariant. The literature was examined to identify strategies (e.g., stakeholder engagement, advisory boards or review panels, multi-group confirmatory factor analysis) for how these two fields establish cultural responsiveness and invariance across diverse groups. The following questions guided the review:

- How do culturally responsive evaluators navigate the use of standardized measures with diverse populations?
- How do evaluators determine if surveys are culturally responsive?
- In what ways do evaluators attend to invariance?
- In the measurement literature, how is culture addressed or included in the analysis of invariance?

Review of the Literature

Empirical articles from 2000-2019 were selected for study one. Articles included evaluation and measurement studies with quantitative or mixed-method approaches, case studies, as well as reflections on past work. Search terms for the critical literature review included “culture,” “cultural,” “cross-cultural,” "culturally responsive," “bias,” "fairness,"

and “equivalence.” Inclusion criteria for articles considered in study one included detailed attention to cultural identities or cultural groups, two or more of the search terms listed above, detailed descriptions of data collection or analysis with diverse populations, and discussions of cultural responsiveness or equivalence. An initial literature search yielded 204 articles. Upon further examination, 53 articles met all of the selection criteria and were thus included in this review. Of the articles included, 31 focused on measurement and 22 focused on culturally responsive evaluation. Most articles came from the *American Journal of Community Psychology*, *American Journal of Evaluation, American Journal of Preventive Medicine*, *Applied Measurement in Education, Canadian Journal of Program Evaluation, Educational and Psychological Measurement, International Journal of Testing, Measurement and Evaluation in Counseling and Development, and New Directions for Evaluation*. Populations included in the evaluations spanned the globe including, but not limited to, Sub-Saharan Africa, Native American and Alaskan Natives, Indigenous and non-Indigenous Australians, and *quilombos* in Brazil. Programmatic context in the culturally responsive evaluation literature covered international, indigenous, and Western contexts. A range of program context were represented in this sample of articles including health (e.g. violence prevention programs, HIV/AIDs awareness and prevention, well-being, sobriety), education (e.g., secondary and post-secondary), as well as family/community/cultural values (e.g. strengthening family relationships and connections to cultural values).

Review Strategy and Analysis

Articles which met the inclusion criteria were read repeatedly, analyzed and summarized in two ways. First, I wrote a one-page summary of notes for each article to allow for an in-depth, rich qualitative analysis. Second, each article was cataloged as a row of data in a spreadsheet. Included in the summary table were descriptions of the program or study, descriptions of the study sample (e.g., discussions of identities, cultures), data collection tools and any discussion of responsiveness and/or equivalence. Data was analyzed using a constant comparative approach (Corbin & Strauss, 2014). In a constant comparative approach, data are analyzed through constant comparisons where one small piece of data is compared for similarities and differences, data similar to one another are grouped together under a theme and the emergence of multiple themes provides a foundation for drawing conclusions (Corbin & Strauss, 2014). Once the notes and summary matrix was completed, I synthesized and analyzed the data for themes and patterns across studies. The themes were used to frame an understanding of the strategies used across these two fields and to identify core considerations for imagining quantitative surveys as culturally responsive. With the questions above guiding my analysis and synthesis of the empirical literature, I identified five themes that reflect my guiding questions: 1) inclusion of cultural experts, 2) establishing cultural relevance, 3) questioning the validity of standardized measures, 4) the use of multiple methods, and 5) inclusion of culture in discussions.

Descriptive Analysis

Inclusion of cultural experts

In both culturally responsive evaluation and measurement studies, cultural experts were called in to support the review, revision, and analysis of quantitative tools used in cultural populations. Culture in items is about the construct as well as the operationalization of the construct in daily life, in the lived context of the stakeholders. Stakeholders in the populations where the evaluations or tests occurred were often called upon to serve as cultural experts. Cultural experts served as advisory council members, consultants for survey review, and critical examiners of problematic items found using statistical techniques.

Stakeholders who served as cultural experts reviewed tests and items, suggested revisions of problematic items, highlighted items which were difficult to understand, flagged items which may be influenced by cultural experiences, commented on word choice, and identified items which may be contextually irrelevant, culturally inappropriate, illegal, or insensitive (Allen et al., 2014; Alkon et al., 2001; Bowen & Tillman, 2015; Coppens et al., 2006; Jenzen et al., 2015; LaPoint & Jacksons, 2004; Mohatt et al., 2004). Additionally, stakeholders provided clarity and face validity (Bowen & Tillman, 2015; Jenzen et al., 2015). LaPoint and Jackson (2004) included students, teachers, staff, family, and community members as stakeholders in their evaluation of a program focused on Black students in low-income urban high schools. According to LaPoint & Jackson (2004), stakeholders “reviewed instruments by assisting evaluators in developing, selecting, or refining items or questions for surveys, focus groups, and

interviews…evaluators were concerned that the content of information was contextually and culturally responsive to participants' experiences in interventions" (p. 30). Engaging stakeholders as cultural experts also led to the clarification of core constructs in strengths oriented ways. For example, in working with Indigenous populations, Allen et al., (2014) wrote, "the term sobriety as used here is a locally defined, indigenous concept that includes abstinence and nonproblem drinking, as well as recovery from alcohol abuse, as well as broader components of well-being" (p. 126). Overall, cultural experts provided critical perspective on the efficacy of quantitative measures.

After a survey is launched, analysis may reveal a subset of items as biased against a particular group. Studies examining items which function differently across cultural populations (differential item functioning) often relied on a panel of subject matter experts or individuals who have expertise in the cultural groups of interest to determine what may have caused the item to be biased. Both statistical techniques and judgmental reviews by content and cultural experts were used to evaluate items displaying differential item functioning (Canel-Cinarbas et al., 2011; Elosua & Lopez-Jauregui, 2007; He & Wolfe, 2010; Huang et al., 2016; Maddox et al., 2015). When statistical analysis flags items for differential item functioning (DIF) a common next step is to submit the items for judgmental review to identify potential causes of DIF (He & Wolfe, 2015). Though it is not uncommon to bring in experts to review the data for potential causes of DIF, these experts are rarely drawn from the same group for which the test or survey was designed. Also rare, are discussion of the sociocultural biases of panelists which may influence findings as to why items functioned differentially across groups.

One study included panelists from the sampled population as cultural experts (Huang et al., 2016) and another study discussed the potential implicit biases of cultural experts serving as panelist (Ross & Okabe, 2006). In examining what caused DIF across U.S. and Mainland China samples of students, Huang, Wilson, & Wang (2016) asked bilingual experts as well as 15 year-old students who were the target audience for the test to review the problematic items and provide feedback. Ross and Okabe (2006) found that the Japanese, male experts who reviewed items flagged for DIF consistently noted items as biased for women. The panelists justifications were shaped by implicit biases related to the perceived capacity of Japanese women to excel in the content area of the items (e.g., language development).

Establishing cultural relevance

A number of studies focused on the importance of cultural relevance of quantitative tools used with diverse populations. Cultural relevance was discussed at the item and construct levels as topics or questions which may not mean the same thing in different cultures, may not fully capture the intended construct across cultures, or may be culturally insensitive. Sy et al. (2015) developed a rubric to evaluate the growth of resilience in Hawaiian adolescents and included culturally relevant practices of Native Hawaiian lifestyles and customs, folklore, and language proficiency. In contrast, when working in Latino communities, Clayson et al. (2002) found the constructs of civic engagement and self-sufficiency were problematic. In reflecting on their collaborative work to define and measure outcomes, Clayson et al. (2002) wrote, "while the concept of 'self-sufficiency' is a positive value indicating individual responsibility within the

dominant European-American paradigm, to those from the Latino Diaspora, the general concept has an inter-dependency component and includes la familia and the broader community" (p. 39). The measure was reflective of Western norms and values. Similarly, Botcheva et al. (2009) found that the concept of 'choice' on their measure was "a particularly Western concept and does not characterize the experiences of people in Zimbabwe" (p. 183). Cultural relevance and conceptual understanding, one's ability to accurately respond to the items, constructs, and response options can impact invariance and can also reinforce a set of dominant values irrelevant to the participants.

Cultural relevance also had a contextual component. Examples in the literature spoke to the importance of cultural understanding in order to evaluate the cultural relevance of an instrument. For example, when working with Indigenous populations on a tobacco intervention program, Unger et al. (2008) wrote,

> If a survey merely asks adolescents whether they use tobacco, respondents may be unsure about whether they should report only their commercial tobacco use or their ceremonial tobacco use as well. A survey with ambiguous questions is likely to yield uninterpretable results. (p. 139)

Similarly, when attempting to measure the construct of 'household assets' with *quilombos* in Brazil, Bowen and Tillman (2015) reflected, "the construct of household assets took time in measuring because many household items related to agriculture were shared across families. Therefore, having a nuanced understanding of *quilombolas* has helped us as we analyzed the data" (p. 34). An understanding of the cultural context within and surrounding the evaluation and measurement tools was an important consideration in establishing cultural relevance.

An important consideration related to cultural relevance is cultural affirmation, or strengths-based approaches rather than deficit based approaches. Culturally affirming measures reflected multiple cultural realities rather than the imposition of a Western framework or Western values, and did not convey stereotypical conventions of a culture or identity (see Allen et al., 2014; Bowen & Tillman, 2015; Coppens et al., 2006; Luyt, 2012; Mohatt et al., 2004; Sy et al., 2015). Allen et al. (2014) used a community approach to develop a culturally anchored measure of protection factors for rural Alaskan Native youth to prevent risk of suicide and alcohol use. They recognized that asking direct questions about trauma, suicidal thoughts, or alcohol use was culturally unacceptable or discomforting. As a result, they chose to adopt and adapt tools which were strengths-based and used positive psychology, stating "we significantly adapted this item pool, emphasizing cultural beliefs and experiences that make life enjoyable, worthwhile, and provide meaning in life, without reference to the presence or absence of suicidal feelings" (Allen et al., 2014, p. 130). In the case of Allen et al. (2014), a strengths-based approach to a culturally sensitive construct was a critical consideration. In contrast, Coppens et al. (2006) experienced multiple challenges in the use of the funder mandated standardized instrument when measuring the impact of a Cambodian Dance program for immigrant youth on developing deep cultural connections with their families. Coppens et al. (2006) received feedback from community members and realized the measure that was “reflective of an individualistic framework evident in American culture, would not be valid in determining the success of a Cambodian program that emphasizes a collectivistic perspective" (p. 325). They went on to explain that items on the measure

conveyed a sense of values rooted in Western and individualistic notions, stating, “there appeared to be a strong inference that being internally oriented was valued more highly than giving importance to others in one's life or to the context of the event” (p. 325). Using a measure which emphasized individualism in a cultural group that values collectivism, in a program designed to help Cambodian youth embrace their heritage while navigating American life and culture was especially antithetical to their purpose. Cultural relevance focused on how constructs such as masculinity (Luyt, 2012), violence prevention (Alkon et al., 2001), life satisfaction and wellbeing (Lau et al. 2015), and stigma (Vogel et al., 2013) manifested within diverse cultural populations. Cultural affirmation focused on how those constructs were rooted, or not rooted, in Western values and experiences.

Questioning the validity of standardized measures

The lack of standardized instruments normed on diverse cultural populations, conversely thought of as the abundance of tools normed on Western or European American populations, emerged as another theme. Empirical articles which discussed both culture and standardized instruments questioned whether the standardized measures functioned well for diverse populations. Authors of these articles recognized the lack of reliable and valid information available for diverse populations and considered issues of equivalence across populations.

Central to this theme is that researchers did not assume standardized measures were culturally responsive and invariant. Those conducting evaluations or evaluating the efficacy of measures in diverse populations noted concerns about the impact of using

tools developed and standardized in "very limited geographical and social setting, namely, white middle-class North American and European children and families" (Abubakar & Van de Vijver, 2017, p. 197) to draw accurate conclusions about historically marginalized populations (See Alkon et al., 2001; Ausili et al., 2019; Bowen & Tillman, 2015; Byrne et al., 2007; Cauffman & MacIntosh, 2006; Coppens et al., 2006; Jenzen et al., 2015; Small et al., 2006). The lack of standardized instruments developed and tested with culturally diverse populations was a critical consideration because it undermined the ability of researchers to draw accurate conclusions across populations (e.g., Lakin et al., 2012), accurately capture program outcomes (e.g. Coppens et al., 2006), support diverse communities (e.g., Bowen & Tillman, 2015), or allocate limited resources (e.g., Cauffman & MacIntosh, 2006). These challenges were not mutually exclusive. As Bowen and Tillman (2015) describe,

> These examples raise the question of the representativeness of international standardized measurements for communities in which they may not be applicable or effective. It compromised our research and limited our ability to represent the population that we were studying within their cultural environment as well as to assist them. (p. 35)

Instruments which are not culturally responsive and invariant can create compounding problems, inhibit the collection of accurate and trustworthy data, and impact service delivery with diverse communities.

The lack of available information regarding reliability and validity, or evidence pertaining to linguistic or measurement equivalence for historically marginalized populations, also emerged as a theme (Alkon et al., 2001; Cauffman & MacIntosh, 2006;

Coppens et al., 2006; Prelow et al, 2000; Small et al., 2006). In evaluating a mental health screening tool for use with racially diverse incarcerated youth Cauffman and MacIntosh (2006) noted the lack of psychometric data on reliability and validity for diverse populations and expressed concern that, “this has serious implications for the interpretation of studies that have examined differences in the prevalence of mental disorders for juvenile offenders of various ethnic backgrounds" (p. 503). Equivalence was discussed but not always statistically examined in these articles. Small et al. (2006) used a standardized instrument required by a funder to evaluate a mental health program serving Hmong community members. The authors reflected on challenges with the standardized measure, writing “some questions were perceived by the parents as irrelevant…some of the questions regarding mental health concepts did not have an equivalent translation in the Hmong language" (Small et al., 2006, p. 361). In working to measure the impact of a violence prevention program with children and families across three ethnic groups, Alkon et al. (2001) wrote, “most relevant standardized instruments had been validated with one ethnic group, usually European Americans, and did not have information on conceptual equivalence for other ethnic groups” (p. 51). Across evaluation and measurement literature, studies questioned the efficacy of measures standardized on White populations, empirically tested the invariance of measures used across cultural groups in order to draw sound conclusions, and noted how standardized measures could be problematic in evaluating program outcomes, delivering services, and allocating needed resources.

The use of multiple methods

The use of multiple methods was another theme which emerged in the literature. Multiple methods were used in three primary ways, 1) to offset the use of tools standardized in predominantly Western samples, 2) to capture more robust information related to cultural populations, or 3) to better articulate program impact on cultural populations. Multiple methods were used to establish validity evidence when using tools standardized with predominantly White populations. The use of multiple forms of data collection and data triangulation offset the data collected with standardized measures by providing additional avenues for community members or respondents to share their understanding of program outcomes. Butty et al. (2004) recognized that one of the tools used for their program evaluation with ethnically diverse community members was standardized on a non-ethnically diverse population, questioned the validity of inferences which could be drawn, and addressed validity concerns by adding additional methods of data collection. Some of these methods were still quantitative, but were screened for use "based on the extent to which they were culturally sensitive by means of their form, language, and content" (Butty et al., 2004, p. 44). In another example, Mamaril et al. (2018) integrated interviews and talk story sessions into a program evaluation with Native Hawaiian community members in an effort to track the progress of program participants after realizing the survey alone did not accurately capture the experiences of all participants. The authors reflected, "qualitative tools appear to be more meaningful indicators of success, and coupled with well-chosen quantitative tools, they can capture more accurate portrayals of participants' experience" (Mamaril et al., 2018, p. 49). The

integration of multiple methods provided additional information to contextualize the use of standardized instruments and explain why such measures may be ineffective.

Qualitative methods were used to capture more robust information regarding how standardized tools functioned in diverse populations. Psaki et al. (2016) used multiple methods when conducting research on mental health with community members in low-income settings to evaluate if a diagnostic tool developed using Western criteria was effective for people in Kumasi Ghana. They wrote, “by supplementing common quantitative approaches to scale validation with qualitative data analysis, we highlight the shortcomings of limiting scale adaptation to quantitative analyses when conducting mental health research in non-Western Settings" (Psaki et al., 2016, p. 341). Some researchers used qualitative methods to revise quantitative instruments for measurement equivalence or to more fully identify a construct for a given population (Jenzen et al., 2015; Luyt, 2012). In evaluating a program for inner-city, ethnically diverse students, Jenzen et al. (2015) used multiple methods and found that the construct of hope, as measured by the standardized tool, was too narrowly defined. Jenzen et al., (2015) explained that,

> Qualitative methods suggested that involvement in City Kidz resulted in a construct of hope that was more holistic than the purely cognitive, goal-oriented survey questions based on the agency and pathway thinking dimensions of hope found in the Children’s Hope Scale. (p. 52)

The measure Jenzen et al. (2015) used was standardized on a sample that was demographically different than the students participating in the City Kidz program. In another example, Sy et al. (2015) used multiple methods to validate their findings related

to program impact on Hawaiian youth, writing, “this corroboration between different data types indicates that the measures to evaluate this program’s primary outcomes—students’ understanding of, identification with, and practice of Hawaiian values—were valid and reliable" (Sy et al., 2015, p. 1524). The use of multiple methods allowed practitioners in evaluation and measurement to better understand to what extent standardized tools were accurate and reliable when used with diverse populations.

In the measurement context, when items on standardized measures were flagged for differential item functioning, panels of experts were often consulted to identify potential causes of DIF (Elosua & Lopez-Jauregui, 2007; He & Wolfe, 2010; Huang et al., 2016). Panelist are often consulted to provide a judgmental review of the flagged items in order to offer insight into why an item might function differentially across groups. Huang et al. (2016) explained that cultural familiarity and relevance may have caused DIF in their study, and upon further review by panelist they found that, “the items on the 'grand canyon' favoured US students. In addition, items on 'forest fire', 'genetically modified food' and 'sun screen' contained subjects that seem to be more familiar to US students, and were found to favour US students" (p. 387). Such reviews provided insight into why items may function differently and how cultural familiarity may have played a role in differential functioning.

Multiple methods were also used to determine program outcomes in diverse populations (e.g., Botcheva et al., 2009; Coppens et al., 2006; Jenzen et al., 2015; Sy et al., 2015). When evaluating a HIV/AIDS program in Zimbabwe, Botcheva et al. (2009) used a standardized survey and discovered students had written poems as a part of a

homework assignment. The authors decided to include the poems as a form of qualitative data as they felt the poems reflected student experiences in the students' own voices and were more culturally resonant than the survey. As the authors explained, "the importance of narrative within the Zimbabwean culture and the importance of artistry within the youth culture made the poetry an ideal data source for measuring the project outcomes" (Botcheva et al., 2009, p. 184). The addition of data from the poems provided a more robust picture of program outcomes with the Zimbabwean students. While the survey results showed little change in HIV/AIDS prevention, the poems did, leading the authors to believe that although the survey was used to measure self-efficacy in South Africa, it did not reflect preteen Zimbabwean culture (Botcheva et al., 2009). In another example, Coppens et al. (2006) encountered a number of challenges in the use of their standardized instrument but felt the addition of qualitative methods helped them document the successful attainment of program outcomes and, "portrayed the richness of the Cambodian culture and the uniqueness of our program in contrast to the quantitative data required by the MCC that focused on individualistic perspectives" (Coppens et al., 2006, p. 329). From an analysis of the empirical literature, qualitative methods can be used in addition to standardized measures to capture a more robust understanding of program outcomes and cultural populations.

Inclusion of culture in discussions

Multiple studies adopted or developed frameworks grounded in culture to guide their analysis and make meaning out of the findings. Articles in this study used the literature to establish cultural frameworks for conducting analysis, linked their findings

back to cultural attributes, or adopted existing cultural ways of understanding. Researchers used frameworks to identify cultural characteristics of the population included in the study, such as individualism or collectivism (e.g., Asil & Brown, 2016; Carrola et al., 2012), or to seek out additional understandings for findings across cultural groups (e.g., Kornilov et al., 2016). Several articles discussed using literature to establish frameworks for conducting their research and examining instruments from a cultural perspective (Abubakar & Van de Vijver, 2016; Alkon et al., 2001; Banks, 2006, 2012; Botcheva et al., 2009). As an example from the measurement literature, Banks (2006, 2012) worked to provide a definition of culture and to establish a framework for evaluating and hypothesizing cultural bias in education testing prior to empirically testing the hypothesis. Banks (2012) developed a seven-step process for identifying if inferential reading items were more prone to cultural bias than literal reading items among a group of Hispanic, Black, and White students. Banks (2006, 2012) used literature to identify and describe general cultural characteristics as a framework for examining items for potential bias prior to empirically testing for differential item functioning, differential bundle functioning, and differential distractor functioning. The author hypothesized that distractors, which were culturally relevant for one group but not another, may play a role in whether students get the items correct or incorrect.

Examples emerged specific to work with Indigenous populations and the importance of prioritizing Indigenous approaches to research (Crooks et al., 2018; Mohatt et al., 2004). Crooks et al. (2018) explained, "our team embarked on this evaluation with a commitment to a two-eyed seeing approach to data collection and interpretation…" (p.

461). More concretely, adopting this framework meant they reflected on the context of the evaluation, prioritized respectful relationships and reciprocity, ensured their work was of benefit to the community, and shared knowledge in ways that were culturally appropriate (Crooks et al., 2018). Grounding research in cultural frameworks shaped investigations of how measures were culturally appropriate and how findings were interpreted in a cultural context.

Discussion

In this study, I examined two distinct bodies of literature simultaneously, culturally responsive evaluation and measurement across cultural groups, to derive considerations for rendering quantitative measures culturally responsive and invariant. In this critical examination of the literature, I identified five themes reflecting my guiding questions: 1) inclusion of cultural experts, 2) establishing cultural relevance, 3) questioning the validity of standardized measures, 4) the use of multiple methods, and 5) inclusion of culture in discussions. In this section, I provide a discussion of these themes and describe how each one relates back to culturally responsive evaluation and measurement research. Although discussed separately, these themes are in fact interconnected, so I explore the intersections between themes further in this section. Finally, I present critical considerations along with guiding questions for evaluating standardized measures as culturally responsive.

The inclusion of cultural experts as a finding in this study was not surprising, but when and which cultural experts were involved presented an interesting contrast between culturally responsive evaluation and measurement approaches. Engaging community

stakeholders is a fundamental element of culturally responsive evaluation (Frierson et al., 2002; Hood, et al., 2015; Hopson, 2009). Cultural experts helped to guide the evaluation process, examine items for cultural relevance and face validity, or identify potential reasons an item might function differently across populations. Evaluators using culturally responsive and participatory approaches often worked with cultural experts to examine measures prior to using them to evaluate their programs. Additionally, once a measure was in use, culturally responsive evaluators listened to feedback from community members regarding the efficacy of these instruments. As a result, issues were identified throughout the evaluation rather than at the end, when data were collected and analyzed. In contrast, the measurement articles included often examined equivalence and brought in cultural experts to critically examine items that had statistically been flagged for bias. As an additional contrast, culturally responsive evaluators often worked directly with the populations engaged in the measurement, whereas measurement practitioners brought in external cultural experts, typically not directly involved in responding to the standardized instrument.

Further, considerations of equivalence and the validity of standardized tools used across diverse populations seemed to happen at two different points in time in the evaluation or measurement process and involved cultural experts in different ways. Multiple studies include discussions of culture in shaping the study or in providing context to the findings. Of note is that in the invariance studies stemming from the measurement literature, there was no mention of confirming the findings in the discussion related to culture with cultural stakeholders. In culturally responsive evaluation, engaging

stakeholders in data analysis and reporting is strongly recommended (Frierson et al., 2002; Hood, et al., 2015). Inviting cultural stakeholders' feedback serves as an opportunity to attend to multicultural validity threats related to voice and power (Kirkhart, 2013), correct any researcher bias that may have influenced the conclusions drawn, and correct any deficit orientated conclusions that could cause harm to the population studied.

Establishing cultural relevance was an important consideration in both measurement and evaluation. Cultural relevance, multiple methods, and questioning data validity are interconnected. In this study cultural relevance included making sure items and constructs were understood as intended across cultures. Multiple methods of data collection were used to supplement standardized measures identified as culturally irrelevant or invariant. Hood et al. (2015) as well as Mertens (2007), argue for the importance of including multiple methods in culturally responsive evaluation. Methodology and methods are critical considerations in CRE as choice of method impacts how culture is, or is not, represented in the data. Standardized methods and Western researchers have a history of harm in culturally diverse communities (Cram, 2016; Crooks et al., 2018). Multiple methods of data collection were used to capture a more robust understanding of the cultural population, the contextual factors influencing the program, or the ways in which diverse communities may respond to a measure. Emergent in this intersection is the idea that cultural insensitivity and the imposition of Western values are threats to multicultural validity. Kirkhart (1995) discusses methodological validity as an aspect of multicultural validity and argues that

measurement issues related to culture are centered on relevance and equivalence. In addition, Kirkhart (2013) argues that the imposition of dominant epistemologies on historically marginalized groups is a threat to validity.

In measurement, cultural relevance and validity evidence were established using statistical techniques and review panels, but the value and use of multiple methods was largely undiscussed in the set of articles examined. However, studies in the measurement literature have referenced the need to examine not only what is happening (using measurement techniques) but also why these phenomenon are happening, and they specify the need for qualitative approaches in conjunction with quantitative approaches (e.g., Allalouf, Hambleton, Sireci, 1999; Byrne & van de Vijver, 2010, Grover & Ercikan, 2017; Kato, Moen, & Thurlow, 2009).

Researcher positionality emerged as another contrast related to establishing cultural relevance. Evaluator positionality and bias is discussed in CRE through topics of positionality, reflexivity, and cultural competence (AEA, 2011; Gordon, et al., 1990; Hopson, 2003; SenGupta et al., 2004; Symonette, 2004). Authors in culturally responsive evaluation explicitly discussed the importance of having culturally sensitive members, or researchers from the same cultural group, as the stakeholders and program participants, and the challenges and limitations which present when only involving cultural 'outsiders' (Bowen & Tillman, 2015; Clayson et al., 2002; Crooks et al., 2018; LaPoint & Jackson, 2004). Abubakar and Van de Vijver (2016) spoke to culturally decentering test in the development process so that items and content are not differentially familiar, but discussions of researcher positionality and bias in data analysis were largely unexamined.

Decentering speaks to the instrument itself, but not to the cultural competence of those who analyze and draw meaning from the instruments. There is no evidence from the measurement articles included in this sample to suggest cultural competence is a skill necessary for accurate and trustworthy data analysis and interpretation.

Also emergent at the intersection of cultural relevance, multiple methods, and standardization are Kirkhart's (2013) principles of power and voice as they relate to multicultural validity. Multiple articles critically questioned how and for whom a measure was standardized. Studies examined how the tool was standardized and for which populations, questioning if the data collected by a tool standardized in predominantly White, Western populations would produce valid data for more diverse populations. Examining invariance was most clearly demonstrated in articles which sought to establish invariance across cultural groups. The *Standards for Educational and Psychological Testing* (1999) encourage researchers to establish validity evidence when using tests with diverse subgroups of a population. Compounding the concern of how instruments were standardized and how these instruments might function in more diverse populations, was the lack of available reliability and validity evidence for specific cultural populations. Articles in culturally responsive evaluation questioned equivalence, but rarely provide empirical evidence of equivalence. Their questions of equivalence stemmed from discussions with their cultural advisory councils or in conversations with community members who had responded to the instruments and provided feedback on concepts that were unfamiliar, words that did not have an equivalent meaning, or culturally irrelevant items (e.g., an item which referenced a boat used in a rural location

within a landlocked country). A core principle of culturally responsive evaluation practice and multicultural validity is attention to power (Hood et al., 2015; Hopson, 2003; Hopson, 2009; Kirkhart, 2013; Mertens, 2011). The continued approach of standardizing measures on predominantly White populations perpetuates ethnocentrism in research (Gordon et al., 1990). Examining which populations were included to standardize a measure pushes back on the perpetual prioritization of Western research interests and works to shift power dynamics by advocating for the inclusion of more diverse representation and ways of knowing in standardized measures. The lack of available instruments standardized on populations beyond Western cultures, in addition to the lack of reliability and validity evidence for cultural populations, serves to further marginalize and silence populations beyond Western cultural groups. Questioning how and for whom a tool was standardized is a critical aspect of culturally anchored methodology, disrupts ethnocentric approaches to research, and calls attention to issues of power and voice for marginalized populations.

Critical Considerations

From the discussion, I have generated two critical considerations (which I supplement with guiding questions) for examining to what extent standardized measures are culturally responsive. These considerations include attention to voice and diverse lived experiences as well as ways and methods of establishing cultural relevance and invariance. This section outlines these two considerations, followed by guiding questions.

Attention to voice

Values and cultural ways of knowing are embedded in instruments used to measure underlying constructs. Prior to distributing a standardized instrument, considering which values are reflected in the questions and constructs on the instrument can provide one avenue for determining if a tool is culturally responsive. One possible way to consider voice is to create a cultural council of stakeholders directly impacted by the outcome of the measure or who are the target population for responding to the measure. This group can provide considerable perspective on if the instrument is culturally relevant, is reflective of multiple values systems, and is deficit or strengths-based. This council can also examine the measure for cultural sensitivity, language, content, and culturally relevant or irrelevant context. Additionally, a cultural council can provide feedback on data analysis, interpretation, and reports. Developing a cultural council to critically examine quantitative measures addresses methodological threats to multicultural validity by allowing researchers to identify if the measures are culturally responsive, establish language equivalence, and identify if there is values incongruence between what the measure is asking and the cultural values of the stakeholders. A cultural council can also provide perspective on whether a measure accurately reflects the lived experiences of stakeholders. Establishing cultural responsiveness and values congruence creates space for multiple ways of knowing to emerge. Examining existing literature related to the instrument to better understand how multiple perspectives were included in the development, testing, and validation process may serve as an means to attend to voice.

Attention to voice involves the voice and experiences of stakeholders as well as the voice and positionality of the researcher. As Frierson et al. (2010) write, “data do not speak for themselves nor are they self-evident; rather, they are given voice by those who interpret them” (p. 91). The framework used to conduct data analysis and findings is that of the researcher. Researcher positionality and bias may influence the interpretation and reporting of findings; as such, the cultural competence and reflexivity of the researcher in how their voice shapes determinations of measures as culturally responsive is critical. Who was included in the development and testing of the measure? Are questions on the tool reflective of multiple ways of knowing or reflective of diverse lived experiences? Will answering questions from a cultural perspective reflect negatively on the respondent in any way (e.g. ceremonial tobacco use)? In higher education, what student development theory supported the development of the constructs measured? Whose voice and experiences will be amplified through data collected by this instrument? Whose voices and experiences will be minimized?

Establish cultural relevance and invariance

Qualitative approaches such as focus groups, life histories, ethnographies, and interviews used in combination with standardized measures, or to evaluate standardized measures, can provide a more robust reflection of the cultural context, the culture of the stakeholders, and the findings. Establishing measurement invariance before making cross-group comparisons of student outcome achievement is critical to ensuring the measure functions without bias across groups. Establishing invariance provides evidence that the instrument is fair for diverse populations and accurate conclusions can be drawn.

Is there research available on the background of how the measure was developed? Is there reliability or validity evidence for the population of interest? To what extent is a measure invariant across populations?

A critical examination of the literature across two distinct professions, culturally responsive evaluation and measurement, lead to the development of five key themes. In positioning these five themes in the broader landscape of research and theory related to cultural responsiveness and invariance, two core considerations emerged for practitioners which address voice, lived experiences, cultural relevance, and invariance. Culturally responsive evaluation and multicultural validity shaped the lens through which I conducted this study. Kirkhart (2013) identified five threats and justifications for multicultural validity: methodological, experiential, relational, theoretical, and consequential. In attending to the threats and justifications, Kirkhart (2013) also provides a checklist which considers history, location, power, voice, relationships, time, reciprocity, plasticity and reflexivity in the argument for validity while moving through an evaluation process. The emergent core considerations align with many of these threats and justification of multicultural validity, begging the following questions for future consideration: In what other ways are their overlaps between the emergent themes and core considerations in this study and Kirkhart's (2013) threats and justifications? How, or to what extent, do each of the items on Kirkhart's (2013) checklist relate to the themes and core considerations developed in this study? How can the core considerations outlined in study one support researchers in establishing multicultural validity?

CHAPTER V

RESULTS: STUDY TWO

The purpose of study two was to examine to what extent the National Survey of Student Engagement is invariant for first-generation and continuing-generation college students. Establishing invariance provides evidence of construct validity and suggests that the measure examined is a fair measure for all students.

Sample Demographics

The sample in this study was half first-generation college students (N = 3,000) and half continuing-generation college students (N = 3,000). All students in the sample were at the same institution where they began their post-secondary career; the sample did not include transfer students. All students in the sample were 18 to 23 years old. In this sample, 4.9% of students were international and 94.5% were not. The majority of students were white (41.6%), then Hispanic or Latino (24.8%), Asian, Native Hawaiian, or other Pacific Islander (15%), American Indian, Alaska Native, or Multiracial (10.9%), or Black or African American (4.9%). The majority of students (98.6%) were not veterans and had not been diagnosed with a disability (88.4%). The majority (59.1%) of students identifying as Asian, Native Hawaiian, or other Pacific Islander were also first-generation college students while 49.9% were continuing. For Black or African American students, 48.1% were also first-generation college students while 51.9% were continuing. The majority (83.3%) of Hispanic or Latino students were first generation, while 16.7%

were not. For White students, 28.9% were first generation and 7.1% were not. More than half of American Indian, Alaska Native, or Multiracial students were continuing generation (55.1%), while 44.9% were first-generation.

The student self-reported National Survey of Student Engagement consists of 47 items composing 10 factors. The analyses focused on the influence of multiple latent variables on the observed variables as well as the influence of latent variables on one another. Because the research questions posed involve the influence of latent variables on observed variables and the relationships between latent variables across groups, multi-group confirmatory factor analysis, as a part of the structural equation model family of analysis techniques, was used. Traditional applications of confirmatory factor analysis assume multivariate normality with continuous data and use maximum likelihood estimation. Items on NSSE had Likert-scaled response options which yield categorical data. The sample size obtained for this study was sufficiently large to treat the data as categorical in the MG-CFA analysis using diagonally weighted least squares (WLSMV) estimation, a technique specifically designed for analysis using ordinal data (Li, 2016). Diagonally weighted least squares estimation does not make the same distributional assumptions of multivariate normality required by maximum likelihood estimation, and has been shown to yield less biased and more accurate factor loadings when considering number of categories, sample size, parameter estimates, and standard errors (Li, 2016).

The skewness, kurtosis, and reliability of the items and their scales were examined to identify any possible concerns with the quality of the data even though multivariate normality is not an underlying assumption using the WLSMV estimation

approach. The skewness index (SI) and the kurtosis index (KI) were used to evaluate the measures of skewness and kurtosis. Variables with absolute values of SI > 3.0 are considered highly skewed while variables with absolute values of KI > 10 are considered to be problematic, greater than KI >20 are seriously problematic (Kline, 2011). All estimates of skewness and kurtosis fell within the acceptable parameters for normality for each variable included in the study.

Table 7 summarizes the reliability, mean, and variance for all students, first-generation college students (FGCS), and continuing-generation college student (CGCS) for each of the factors analyzed on NSSE. The calculated reliability statistic is Cronbach's alpha. For each of the factors, the reliability statistic was greater than the common standard of .80 with the exception of the learning strategies factor which had a reliability coefficient of .752 for all students, .754 for first-generation college students, and .750 for continuing-generation college students. Two other factors fell below .80 for continuing-generation colleges students: effective teaching (α = .796) and quality of interactions (α = .777). Given that all values exceeded .70, the measures can still be considered to have acceptable reliability for group level analysis.

Table 7: Reliability, Means, and Variance by Factor and Group

Collaborative Learning (N = 4)	Cronbach's α	Mean	Variance
All	.809	10.74	7.405
FGCS	.802	10.60	7.338
CGCS	.816	10.88	7.435
Reflective and Integrative Learning (N = 7)			
All	.860	19.43	17.279
FGCS	.865	19.44	17.681
CGCS	.855	19.41	16.884

Student Faculty Interactions (N = 4)			
All	.820	7.99	8.136
FGCS	.834	8.10	8.855
CGCS	.803	7.88	7.399
Higher Order Learning (N=4)			
All	.832	11.71	6.709
FGCS	.853	11.69	6.958
CGCS	.811	11.72	6.463
Effective Teaching (N=5)			
All	.820	14.61	9.998
FGCS	.841	14.74	10.951
CGCS	.796	14.48	9.016
Quantitative Reasoning (N=3)			
All	.817	7.17	5.130
FGCS	.825	7.10	5.205
CGCS	.808	7.24	5.045
Discussions with Diverse Others (N=4)			
All	.872	12.12	9.205
FGCS	.890	11.89	10.133
CGCS	.849	12.34	8.191
Learning Strategies (N=3)			
All	.752	8.69	4.136
FGCS	.754	8.67	4.157
CGCS	.750	8.72	4.115
Quality of Interactions (N=5)			
All	.861	25.21	40.048
FGCS	.839	25.51	48.748
CGCS	.777	26.71	35.120
Supportive Environment (N=8)			
All	.889	23.02	28.721
FGCS	.901	22.92	31.685
CGCS	.876	23.12	25.740

To examine invariance, I implemented a series of nested models with increasingly restrictive equality restraints imposed on the parameters. Lavaan (Rosseel, 2012) and semTools (Jorgensen, Pornprasertmanity, Schoemann, & Rosseel, 2020) packages in *R* were used to conduct the analysis. A CFA fitting the 10-factor model to 47 variables for each population of first-generation college students and continuing-generation college

students was estimated separately for each group to establish fit of the factor structure before moving on to subsequent analysis. The first model fit the 10-factor structure to first-generation and second-generation college students simultaneously with no equality constraints to test for configural invariance. Next, the factor loadings were constrained to be equal across groups and the 10-factor model was fit simultaneously to both groups to determine metric invariance. The final model constrained both loadings and thresholds across groups and simultaneously fit the 10-factor model to both populations to determine scalar invariance.

Four goodness of fit statistics were calculated for each of the models to evaluate model fit. The chi-square difference test is significant at $p < .05$, but is highly sensitive to sample size creating the need to examine additional test statistics. The root mean squared error of approximation (RMSEA) $\leq .05$ suggests a well-fitting model with values between .05 and .08 indicating moderate fit. Comparative fit index (CFI) values greater than .95 indicate excellent fit. Standardized root mean squared residual (SRMSR) values $\leq .08$ indicate good fit. I used each of the fit statistics to evaluate the fit of each model (e.g., configural, metric, and scalar) before moving on to the next level of analysis. To evaluate differences and draw comparisons between the nested models, changes in CFI (ΔCFI) were examined with differences greater than .01 signaling a significant change in model fit (Cheung & Rensvold, 2002).

Multi-group Confirmatory Factor Analysis

Table 8 shows a summary of the fit statistics from the confirmatory factor analysis fitting the model identified by Miller et al. (2016) to first-generation college

students and continuing-generation college students separately in order to establish a baseline of model fit prior to the multi-group confirmatory factor analysis. For continuing-generation college students, the model chi-square test was statistically significant ($\chi^2_M(989) = 6920.23,\ p < .05$), indicating the exact-fit hypothesis was rejected. Given the sensitivity of the chi-squired test to large sample sizes, additional fit statistics were examined. The RMSEA was .048, less than .05, suggesting a well-fitting model. The CFI was .98 indicating excellent fit and SRMSR was .044 suggesting good fit. Although the chi-squared test failed, three other fit indices suggest very good fit of this model for this group. The baseline model also fit first-generation college students well. The chi-square model test was statistically significant ($\chi^2_M(989) = 6879.55,\ p < .05$). The RMSEA was .041 indicating a close-fitting model. The CFI was .98 indicating excellent fit and the SRMR was .048 showing reasonably good fit. Although the chi-squared test was failed, the three other fit indices suggest very good fit of this model. With a baseline of model fit established for each group separately, the next step was to examine both groups simultaneously in the multi-group confirmatory factor analysis.

Table 8: Fit Indices for the Baseline Model

Model	X^2	**RMSEA**	**CFI**	**SRMR**
CGCS	6920.23	.048	.98	.044
FGCS	6879.55	.041	.987	.048

Configural Results

Table 9 contains a summary of the fit statistics for the three nested hierarchical models. In the configural model, the factor structure was specified using the same model specified by Miller et al. (2016) and the parameters were freely estimated. The model chi-square test was statistically significant ($\chi^2_{\mathrm{M}}(1978) = 13790.78,\ p < .05$) and the exact-fit hypothesis was rejected. The RMSEA was .048 suggesting a close-fitting model and the CFI was .984, indicating excellent fit. The SRMR was .042 suggesting good fit. Although the chi-squared test was significant, the three additional fit indices suggested very good fit of this model to the data.

Metric Results

In the metric model, the factor loadings were constrained to be equal across both groups. The model chi-square test was significant ($\chi^2_M(2015) = 13988.12,\ p < .05$). The RMSEA was .048 indicating a close-fitting model. The CFI was .984 indicating excellent fit. The SRMSR was .043 indicating good fit. Taken together, the fit indices suggest very good fit of this model to the data. An additional statistic was provided for this level of analysis: the likelihood ratio test statistic. This measure compares the fit of the configural model to the metric model. The chi-squared difference test between the configural model and the metric model was significant ($\chi^2_D(37) = 69.425,\ p < .05$). Given large sample size, the chi-square statistic may be too sensitive. Changes in CFI (ΔCFI) were also examined with differences greater than .01 signaling a significant change in model fit; there was no difference in CFI.

Scalar Results

The scalar model constrained both loadings and intercepts across groups. The model chi-square was statistically significant ($\chi^2_M(2119) = 14252.86,\ p < .05$). The RMSEA was .047 suggesting a close-fitting model. The CFI was .984 indicating excellent fit and the SRMR was .042, an indication of reasonably good fit. Taken together, the fit statistics suggested good fit of this constrained model to the data. The likelihood ratio test statistic was used to compare the fit of the metric model to the scalar model. The chi-squared difference test was significant ($\chi^2_D(104) = 373.44,\ p < .05$). Small changes in RMSEA, CFI, and SRMR suggest scalar invariance was a reasonable conclusion. Changes in CFI (ΔCFI) were also examined, there was minimal difference in CFI.

Table 9: Fit Indices for Configural, Metric, and Scalar Models

Model	X^2	***df***	**RMSEA**	**CFI**	**SRMR**	**Likelihood Ratio Test**
Configural	13790.78	1978	.048	.984	0.042	
Metric	13988.12	2015	.048	.984	0.043	$X^2(2119) = 4252.86$, $p < .05$
Scalar	14252.86	2119	.047	.984	0.042	$X^2(104) = 373.44$, $p < .05$

Unstandardized and standardized factor loadings were estimated for each of the three models (configural, metric, and scalar). Results for all three models are provided in Appendix A. In reviewing the output, .7 was used as a generally acceptable threshold for examining standardized loadings. In the configural model, the unstandardized loadings for first-generation college students fell between .929 and 1.208 for all factors. For

continuing-generation college students the unstandardized loadings fell between .884 and 1.222. The range of standardized loadings was between .685 and .909 for first-generation college students. Two items fell below the .7 threshold: asking for help (.687) and quality of interactions with other students (.685). These items loaded on separate factors of collaborative learning and quality of interactions. Collaborative learning had four items, three of which had standardized loadings between .747 and .828. The quality of interaction factor was composed of five total items, four of which had standardized loadings between .735 and .827. The standardized loadings for continuing-generation college students fell between .631 and .913. Six items fell below the standardized loading threshold for continuing-generation college students. Four of these items were on the same factor, quality of interactions. The items were quality of interaction with students (.631), advisors (.651), staff (.681), and administrators (.665). The factor reflective and integrative learning had seven items, one of which was below the threshold: the integration of diverse perspectives in discussions and assignments (.689). The factor of supportive environment had eight items, one fell below .7: supportive environment in helping to manage non-academic responsibilities (.693). Given that the lowest standardized loadings were still quite close to .7, all indicators were deemed to be adequate for measuring the underlying factors.

In the metric model, the range of unstandardized loadings (constrained to be equivalent in both groups) was from .912 to 1.232 for all factors. For first-generation students, the range of standardized loadings was from .685 to .909. Two items had standardized loadings of less than .7: asking for help (.698) and quality of interactions

with other students (.685). These items fell on separate factors of collaborative learning and quality of interactions. For continuing-generation students, standardized loadings ranged from .631 and .888. Four items had standardized loadings below .7 for continuing-generation college students, all of which were on the quality of interactions factor. Quality of interaction with students (.631), advisors (.674), staff (.679), and administrators (.672) were also below the threshold. Given that the lowest standardized loading were still quite close to .7, all indicators were deemed to be adequate for measuring the underlying factors.

In the scalar model, the unstandardized loadings (constrained to be equivalent across groups) were between .927 and 1.2 for all factors. The range of standardized loadings fell between .688 and .91. The same two items from prior models had standardized loadings of less than .7: asking for help (.688) and quality of interactions with other students (.688). For continuing-generation college students, standardized loadings in this model ranged from .628 to .911. Six items on three factors were below the threshold, four of these items were on the same factor, quality interactions. These are the same items with standardized loadings of below .7 found in the configural and metric models for continuing-generation college students: quality of interaction with students (.628), advisors (.652), staff (.682), and administrators (.659). The factor reflective and integrative learning had seven items, one of which fell below .7: the integration of diverse perspectives in discussions and assignments (.692). The factor supportive environment had eight items, one of which fell below .7: helping to manage non-academic responsibilities (.695). Although below the threshold, each of these loadings was not

sufficiently low enough to cause concern. Table 10 provides summary of the standardized loadings falling below the threshold in at least one of the two groups examined.

Table 10: Summary of Standardized Loadings Below .7 In At Least One Group

	Configural		**Metric**		**Scalar**	
	FGCS	CGCS	FGCS	CGCS	FGCS	CGCS
RI diverse	.742	.689	.731	.705	.740	.693
Asking for help	.687	.725	.698	.711	.688	.733
QI students	.685	.631	.685	.631	.688	.628
QI advisor	.747	.651	.731	.674	.746	.652
QI staff	.735	.681	.736	.679	.733	.682
QI admin	.734	.665	.729	.672	.740	.659
SE non academic	.748	.693	.741	.702	.747	.695

This analysis fit the NSSE 10 factor model identified by Miller et al. (2016) separately to first-generation college students then continuing-generation college students to establish a baseline model and confirm the factor structure for both populations. After confirming the factor structure, both groups were analyzed simultaneously using a multi-group confirmatory factor analysis with an increasingly restricted series of models to establish configural, metric, and scalar invariance. The results of the configural, metric, and scalar invariance models and the comparisons of fit between these models suggested that the National Survey of Student Engagement is invariant for both first-generation and continuing-generation college students. This finding is significant for two reasons. First, the findings validate meaningful comparisons of means across groups. Second, the analysis establishes NSSE as a fair, and unbiased, measure of engagement outcomes for both continuing-generation and first-generation college students.

Ancillary Findings

Examining the latent mean scores across first-generation and continuing-generation college students was not a part of this research study; however, because scalar invariance held, I was able to compare latent means across groups. For this study, the reference group was first-generation college students and the focal group was continuing-generation college students. The standardized factor mean scores (also the effect sizes in this case) for each factor for continuing-generation college students are provided below in Table 11. In a standardized solution, an absolute value greater than two indicates a significant result, significant results are shown in the *p*-value column. In multi-group CFA, the standardized latent means of the second group can be interpreted as an effect size. For ease of discussion the term effect size will be used in Table 11 and in further discussion.

Table 11: Effect Size for Each Factor

Factor	**Effect Size (Standardized Mean)**	***p* value**
Reflective and Integrative Learning	-.02	.46
Higher Order Learning	<.01	.99
Quantitative Reasoning	.07	.03
Learning Strategies	.03	.44
Collaborative Learning	.11	< .01
Discussions with Diverse Others	.13	< .01
Student Faculty Interactions	-.06	.09
Effective Teaching Practices	-.13	< .01
Quality Interactions	.18	< .01
Supportive Environment	< .01	.86

In this case, continuing-generation college student means scores were significantly higher than first-generation college student scores for quantitative reasoning, collaborative learning, discussions with diverse others, and quality of interactions. First-generation college student mean scores were significantly higher than continuing-generation college students for effective teaching practices with a small effect size (d = -.13). The effect size serves as an indicator of the magnitude of impact of first-generation or continuing-generation college student status on a given factor. Effect sizes of .20 or less are considered to be trivial, .40 moderate, and .60 or higher considered to be substantial. Small (d < .20) significant effect sizes were found for five out of 10 factors and should be interpreted with caution. The largest of the effect sizes was for quality of interactions (d = .180).

Discussion

The National Survey of Student Engagement has been established for over 20 years, but little is known about how this instrument functions across diverse populations nor if it is a culturally responsive measure. The purpose of this study was to determine to what extent invariance could be established using multi-group confirmatory factor analysis. In this section, I summarize the findings from study two, position these findings in the broader context of higher education, and discuss to what extent multi-group confirmatory factor analysis extends cross-cultural understanding in this case example.

This study found clear support for the structural and measurement invariance validity of data collected from first-generation and continuing-generation college students. From the configural analysis, higher education practitioners know that first-

generation and continuing-generation college students use the same cognitive framework for responding to items on the survey, or said another way, associate the same items with the same underlying factors as continuing-generation college students. Results from the metric level analysis indicate that the weight or relationship between each factor and item are fundamentally the same across both groups; the students used the response options in the same way. Additionally, the scalar analysis showed that the students responded to the items without bias; scalar equivalence held across groups. Taken together these results indicate that, despite social and cultural differences, first-generation and continuing-generation college students share the same foundational view of student engagement in the campus environment as evaluated by the NSSE.

Given prior research on first-generation college students, I was interested in two specific factors of collaborative learning and student faculty interactions. Because scalar invariance was established, the factor means for each group could be compared. The collaborative learning factor asks students to report on their behaviors including asking other students for help understanding course material, explaining content to other students, preparing for exams by discussing/working through course material with other students and working with other students on assignments. Prior research has shown that first-generation college students feel uncomfortable when they view themselves as struggling and their peers as being successful (Means & Pyne, 2017). Given findings from study two, the finding that continuing-generation college student scores were higher in this area is not surprising. Prior research on first-generation college student has focused on their perceptions and challenges related to interacting with faculty (Means &

Pyne, 2017). The NSSE factor on student-faculty interactions asks students to report on if they talked to faculty about career plans, work with faculty on activities other than coursework, discussed course content with faculty outside of class, or discussed their academic performance with faculty. Contrary to Means and Pyne's (2017) findings, this study shows that continuing-generation college students reported lower scores related to student and faculty interactions than first-generation college students.

Educators and policy makers in higher education are heavily vested in the advancement of attainment outcomes across diverse populations. Prior to graduation, students engage in a number of activities in the curricular and co-curricular environments which support their development and progression towards degree completion. To accurately report outcome attainment, allocate resources, and develop interventions, data collected using outcome measures needs to be trustworthy and accurate across cultural populations. Prominent outcome measures are widely used as indictors of quality in higher education. Data from outcome measures inform interventions for students, and are largely built on theory developed with White samples. As such, outcome measures like NSSE should only be used after examined for validity and reliability across cultural groups. First-generation college students are a heterogeneous group brining into higher education unique lived experiences and often challenges. In a context of perpetually reduced resources and the continued calls for accountability and equity in higher education, data disaggregation and the ability to draw comparisons across groups has become increasingly important. Prior to drawing comparisons across group, invariance should be established. In study two, the invariance of factor loadings and thresholds

across first-generation and continuing-generation college students was established. The results from the multi-group confirmatory factor analysis provide evidence that the National Survey of Student Engagement functions well across both groups and comparisons are appropriate.

Beyond establishing invariance across groups, one of the questions posed in this study was to what extent multi-group confirmatory factor analysis as a statistical approach advances cross cultural understanding. Using multi-group confirmatory factor analysis to establish invariance furthers cross cultural understanding in two ways: by establishing that a survey is measuring the same construct across cultures and by presenting options for within and across group analysis as recommended in the culturally responsive evaluation literature. The quality of factors and factor loadings was not examined in this study but is one example of within and across group comparisons resulting from multi-group CFA that could further cross-cultural understanding.

Culture is a shared set of values and behaviors within a group of individuals which can be characterized by demographic variables and systemic factors such as politics and economics. All individuals develop in contexts shaped by culture, which influences what is taught, how learning occurs, how learning is demonstrated, which ways of learning are considered valid, and the tools used to measure learning (Hughes et al., 1993). Thus, culture may influence measurement equivalence as values, beliefs, and socioeconomic factors influence how people make meaning out of the constructs measured as well as the response options used to measure those constructs. Multi-group confirmatory factor analysis allows researchers to empirically test if constructs have the

same meaning and factor structure across groups, the weight of the relationships between items and factors for each group, and if the scale is used in the same way across groups. When configural, metric, and scalar invariance are established, then differences found across groups are truly due to differences in the construct measured, and not as a result of group membership. When items do not function similarly across groups, the ability to draw valid inferences across marginalized populations is inhibited and may lead to inaccurate conclusions about the construct of study across groups (Zumbo, 1999). Items and constructs that function differently across populations may indicate bias. Banks (2006) explains that "items that are culturally biased have characteristics that are unrelated to the achievement construct being measured but are sensitive for particular cultural groups and affect their performance" (Banks, 2006, p. 115). Establishing invariance yields cross-cultural insights by establishing that two cultures understand a construct in similar ways and confirms that differences are due to the property being measured and not due to cultural bias in the instrument.

Second, using multi-group confirmatory factor analysis to establish invariance facilitates across group and within group comparisons, which is recommended by foundational literature in culturally responsive evaluation (Gordon et al., 1990; Frierson et al., 2010; Hood et al., 2015; Hughes et al., 1993). One of the benefits of multi-group confirmatory factor analysis is that this analysis assigns relative weights between specific observed variables and latent variables that signify the importance of the different variables within and across populations (e.g., Allen et al., 2014). Loadings, or weights, can be compared within each group of interest as well as across groups. In establishing

invariance, the goal is for these loadings to be the same across groups; however, if this level of invariance cannot be determined, useful information is still gained (e.g. Boer et al., 2018; Byrne et al., 2007). Patterns across and within cultures can be examined using multi-group confirmatory factor analysis.

Findings in this study allow me to reasonably conclude that NSSE is invariant at the configural, metric, and scalar level. As a result, I can draw conclusions about first-generation college student experiences using the latent factor means. The results from this study indicate that drawing comparisons between factor mean scores across both first-generation and continuing-generation college students is appropriate and results reflect true difference rather than bias in the measure. Additionally, because invariance was established, I feel more confident that differences between these two groups represent real differences in their engagement in higher education and are not reflective of cultural bias in the instrument. In the context of higher education this is meaningful as one of the core purposes of NSSE is to inform institutional decision-making regarding student outcome achievement.

CHAPTER VI

DISCUSSION AND CONCLUSIONS

I began this research journey by asking questions about how culturally responsive evaluators examine standardized measures for cultural responsiveness, to what extent invariance can be established, and how quantitative measure can contribute to cross-cultural understanding. To explore my questions, I used a case study example of first-generation college student responses to the National Survey of Student Engagement. In the last chapter of this study, I will discuss my tentative conclusions drawn from combining results from study one and study two. As with any good study, oftentimes more questions emerge than answers. As such, I will identify future opportunities for research. Furthermore, I will convey the contributions I feel this work has made to evaluation, measurement, and higher education as well as limitations of this study.

Connections across Study One and Two

From the critical examination of the empirical literature in study one, I identified five themes that reflected my guiding questions: 1) inclusion of cultural experts, 2) establishing cultural relevance, 3) questioning the validity of standardized measures, 4) the use of multiple methods, and 5) inclusion of culture in discussions. Two core considerations emerged from the themes generated by the critical examination of the literature to begin to evaluate if a standardized tool is culturally responsive: attention to voice and establishing cultural relevance and invariance. In this section, these two

considerations are used to examine NSSE as a culturally responsive and invariant measure for first-generation college students.

Attention to Voice

In 2013, researchers launched a revised version of the NSSE instrument. The revision included consultations with campus experts and advisory boards, a literature review, examinations of existing NSSE data, focus groups, cognitive interviews, pilot and psychometric testing (NSSE, 2018). Staff at NSSE conducted two rounds of pilot testing. In the first round, 21,000 students from 19 institutions responded to the survey, 40 students across seven institutions participated in cognitive interviews, and focus groups were hosted at five different campuses (NSSE, 2018). In the second round of the pilot, 50,000 students from 55 institutions responded to the survey, and qualitative information was collected across 12 campuses through 120 cognitive interviews and 10 focus groups. Staff at NSSE facilitated two pilot studies which included both qualitative and quantitative data collection and analysis to determine necessary revisions for the instrument. The sample demographics were combined across both pilots. The demographic characteristics across both samples were two-thirds women, two-thirds White students, and 50 percent first-generation college students (BrckaLorenz et al., 2012). Detailed information on the racial demographics were not provided. Respondents were from predominately White institutions, minority-serving institutions, and religiously affiliated institutions. Demographic data were not provided for the qualitative portions of the pilots (Haeger et al., 2012). The research provided on the revisions of NSSE shows that multiple methods were used, including multiple qualitative methods, to attend to

student experiences and examine how students understand the instrument. Efforts were made to intentionally garner feedback from historically marginalized populations. Through the use of multiple and mixed methods as well as the intentional inclusion of historically marginalized student perspectives, NSSE researchers attended to voice and multiple ways of knowing beyond the default and majority operating perspective of White culture (Hall, 1992; Kirkhart, 2010). Research related to the revisions did not discuss the use of a cultural council to shape the revision process, the positionality and demographics of the researchers were not included.

Cultural Relevance and Measurement Invariance

From study two, I found clear support for the invariance of NSSE for first-generation and continuing-generation college students. The configural analysis shows that first-generation and continuing-generation college students use the same cognitive framework for responding to items on the survey, or said another way, associate the same items with the same underlying factors. The metric level analysis indicates that the weight or relationship between each factor and item are fundamentally the same across both groups; the students used the response options in the same way. Scalar analysis shows that the students responded to the items without bias. Together these results indicate that, despite social and cultural differences, first-generation and continuing-generation college students share the same foundational view of student engagement in higher education as evaluated by NSSE.

The core considerations established from the critical examination of the literature combined with the results from the multi-group confirmatory factor analysis lead me to

conclude that NSSE may be culturally responsive and invariant for first-generation college students in so far as NSSE functions similarly for both first-generation and continuing-generation college students. First-generation college students were included in the samples for the pilot studies to revise NSSE, multiple methods were used to facilitate a comprehensive understanding of how students understand and respond to the items, and scalar invariance held. Specific discussions with a cultural council of first-generation college students would solidify my determination.

Contributions to Research on Evaluation and Higher Education

My study makes four contributions to the research in evaluation, higher education, and measurement. First, I drew on two well established fields to respond to calls for cultural responsiveness in higher education by examining the literature and developing critical considerations, which higher education practitioners can use to begin to evaluate standardized tools as culturally responsive and invariant. The core considerations from study one could serve as building blocks from which to refine and develop a robust set of criteria to evaluate other standardized tools as responsive for historically marginalized populations. This study builds on the critical examination of the culturally responsive evaluation literature conducted by Chouinard and Cousins (2009) and responds to their question of how quantitative approaches can foster cross-cultural understanding. In addition, findings from this study expand on the work of Chouinard and Cousins (2009) by contributing to research on culturally responsive evaluation, specifically how evaluators use standardized measures in practice. Next, I used culturally responsive evaluation and multicultural validity as a lens through which to approach this

study to center first-generation college students in the examination of how a standardized measure, which is often used to make decisions about the quality of student experiences, functions for a diverse population. The current study contributes to the literature in higher education on first-generation college students by supporting the use of NSSE data to make decisions which support first-generation college student success. Prior to the current study, the invariance of NSSE for first-generation college students had not been examined. The current study also makes a contribution of questioning if a long-standing instrument in higher education reflects multiple ways of knowing for first-generation and continuing-generation college students, or if the measure maintained dominant ways of knowing (Gordon et al., 1990; Hughes, et al., 1993; Stanfield, 1999). In addition, the current study responds to calls in the measurement literature to expand research beyond the analysis of countries, racial groups, and gender to groups across socioeconomic statuses.

Limitations

The first limitation to the current study includes capturing students from only one institution type. The sample data provided from NSSE represented students at doctoral-granting institutions with enrollments of 20,000 or more. First-generation college students at smaller institutions, community colleges, or at private institutions may have a different experience, or different opportunities to engage in activities captured by NSSE. Community college students often come to campus with different lived experiences and demographics than students attending mid-size doctoral granting institutions, which may impact how first-generation college students engage in the campus environment. An

additional limitation is that a foundational aspect of cultural responsiveness is stakeholder involvement. Establishing a cultural council to review NSSE for cultural relevance was not implemented as a part of the current study. Using a cultural council would be an area of research for those with interest in first-generation college students and additional historically marginalized populations in higher education not discussed. Another limitation to the current study was the use of first-generation college student status as a proxy variable for first-generation college student experiences in higher education. Though the statistical analysis supported the conclusion that NSSE is invariant across first-generation and continuing-generation college students, given the size of this population in higher education, additional analysis for sub-populations within first generation college students may contribute to a continued understanding of first-generation college student experiences with engagement in higher education. First-generation college students are a heterogeneous group. Given the demographic composition of first-generation students, significant results from the multi-group confirmatory factor analysis may be due to first-generation college student status or another demographic characteristic (e.g., Hispanic, non-Hispanic) where there were larger proportional differences in the sample between first-generation and continuing-generation college students. Finally, a limitation to the current study was the narrow focus on one aspect of validity. The current study did not examine to what extent NSSE could serve as a predictor for either population. Additionally, consequential validity was mentioned, but not examined in the current study. How NSSE is used to make decisions regarding diverse populations could also be of interest.

Future Research: What is Next?

Several recommendations for future research stemmed from the two studies. First, the current study did not examine the strength of relationships or quality of factor loadings. Examining these parameters within and across groups could support institutions in determining how to strategically allocate resources in specific areas that matter to different groups of students. What is the strength of the relationship between the items and constructs across the two populations, specifically for the factors with small significant effects? Another suggestion for future research would be to engage historically underrepresented students beyond focus groups and instead as a cultural council to examine NSSE and other standardized surveys in higher education. How do first-generation college students view NSSE as culturally relevant or irrelevant? Past research has shown the role of familial relationships for first-generation college students, but family engagement and support are not covered on NSSE. Developing a cultural council of first-generation college students to review and discuss the instrument and influences on their engagement not covered by the survey could be enlightening. In addition, to what extent do the core considerations established from study one support the identification of other standardized tools as culturally responsive or not, for populations beyond first-generation students?

Beyond first-generation college students and NSSE, several other areas of future research emerged. Foundational literature in culturally responsive evaluation positions CRE as a mixed-methods approach (Hood et al., 2015), yet empirical literature on CRE that includes the use of quantitative instruments are rare. This study adds to the

knowledge of quantitative methods in CRE, but provoked more questions. To what extent are quantitative or mixed-methods approaches used in culturally responsive evaluation? Foundational literature in culturally responsive evaluation makes recommendations regarding how to develop, implement, and analyze quantitative methods; to what extent are these recommendations followed in the empirical literature? To what extent do culturally responsive evaluators meet, or not, the thresholds established by Kirkhart (1995, 2005, 2013) for multicultural validity? Beyond surveys and standardized measures, to what extent are culturally responsive approaches used to develop tools such as rubrics (e.g. Sy et al., 2015)? I used culturally responsive evaluation and multicultural validity as a lens through which to guide this study. Given the results from study one, future research could examine to what extent are there other areas of overlap between the emergent themes and core considerations in this study and Kirkhart's (2013) threats and justifications. How, or to what extent, do each of the items on Kirkhart's (2013) checklist relate to the themes and core considerations developed in study one? To what extent can the core considerations outlined in study one support researchers in establishing multicultural validity?

In measurement, Banks (2006, 2012) used theory to develop a framework for analyzing cultural bias in tests, to hypothesize items which may be biased, and then empirically test for bias. Given the robust student development theory available in the higher education context, to what extent has the type of framework presented by Banks (2006, 2012) been used to study standardized measures for diverse populations? Drawing on Banks' work in higher education provides a wealth of opportunities to advance

research on how standardized outcome measures perform for historically marginalized populations.

Conclusions

A central focus of this study was addressing the assumption that instruments normed on dominant populations operate and should be used across diverse populations as measures of outcome achievement in higher education without critical examination. This study took questions posed in transformative and culturally responsive evaluation and applied them to standardized measures used with diverse populations in the context of higher education. As this research unfolded, so too did my understanding of the embeddedness of the assumption of how standardized measures function across groups, and the critical role stakeholders play in articulating the relevance of such measures which are used to evaluate student experiences, make resource allocation decisions, and inform policy. The National Survey of Student Engagement was designed as a tool to inform institutional improvement and shape policy decisions without critical consideration for how this tool functioned for those already underserved in higher education. I believe this may be a very clear example of what Kirkhart (1995) refers to as "arrogant complacency." The evidence suggests that the practitioners who develop, implement, and analyze NSSE data have done their due diligence in working to continuously improve the measure; however, institutions in higher education have an obligation to think critically about the tools they use to make claims about student success and how these tools function in an increasingly diverse population. Using culturally responsive evaluation and multicultural validity as lens for approaching the

current research allowed me to more fully understand the ways in which systems and structures amplify some voices, usually those in positions of privilege and power, and systemically silence others. Asking the question, “does this instrument function well, without bias, for all who may take it?” is a simple question with the potential to elevate diverse voices, expose injustices, and provoke transformation.

REFERENCES

A Test of Leadership: Charting the Future of U.S. Higher Education. A Report on the Commission Appointed by Secretary of Education Margaret Spellings. Washington, D.C.; U.S. Department of Education, 2006.

Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational measurement: Issues and practice, 19*(3), 16-26.

Abma, T. A., & Stake, R. E. (2001). Stake's responsive evaluation: Core ideas and evolution. *New directions for evaluation, 2001*(92), 7-22.

Abubakar, A., & Van De Vijver, F. J. (2017). How to adapt tests for sub-Saharan Africa. In *Handbook of Applied Developmental Science in Sub-Saharan Africa* (pp. 197-212). Springer, New York, NY.

AERA, APA, & NCME (Eds.). (2014). Fairness in testing. In *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Alkon, A., Tschann, J.M., Ruane, S.H., Wolff, M., Hittner, A. (2001). A violence-prevention and evaluation project with ethnically diverse populations. *American Journal of Preventive* M*edicine, 20*(1S), 48-55.

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of educational measurement*, *36*(3), 185-198.

Allen, J., Mohatt, G. V., Fok, C. C. T., Henry, D., Burkett, R., & Team, P. A. (2014). A protective factors model for alcohol abuse and suicide prevention among Alaska Native youth. *American journal of community psychology*, *54*(1-2), 125-139.

Allen, M. and Yen, W. (1979). *Introduction to Measurement Theory*. Long Grove, IL: Waveland Press Inc.

American Evaluation Association (2011). *Public Statement on Cultural Competence Evaluation.* Fairhaven, MA. Author: Retrieved from www.eval.org.

Asil, M. and Brown, G. (2016). Comparing OECD PISA reading in English to other languages: Identifying potential sources of non-invariance. *International Journal of Testing*, *16*(1), 71-93.

Astin, A. W. (1984). Student involvement: A developmental theory for higher education. *Journal of college student personnel*, *25*(4), 297-308.

Ausili, D., Barbaranelli, C., & Riegel, B. (2020). Generalizability of the self-care of diabetes inventory across cultures and languages: Italy and the United States. *Evaluation & the Health Professions*, *43*(1), 41-49.

Avery, D. R., Tonidandel, S., Thomas, K. M., Johnson, C. D., & Mack, D. A. (2007). Assessing the Multigroup Ethnic Identity Measure for measurement equivalence across racial and ethnic groups. *Educational and Psychological Measurement*, *67*(5), 877-888.

Banks, K. (2006). A comprehensive framework for evaluating hypotheses about cultural bias in educational testing. *Applied Measurement in Education*, *19*(2), 115-132.

Banks, K. (2012). Are inferential reading items more susceptible to cultural bias than literal reading items? *Applied measurement in education*, *25*(3), 220-245.

Bledsoe, K., & Donaldson, S. I. (2015). Culturally responsive theory-driven evaluation. *Continuing the journey to reposition culture and cultural context in evaluation theory and practice*, 3-28.

Bocock, R. (1992). The cultural formations of modern society. In S. Hall, & B. Gieben (Eds.), *Formations of modernity* (pp. 229-274). Cambridge, GB: Polity Press.

Boer, D., Hanke, K., & He, J. (2018). On detecting systematic measurement error in cross-cultural research: A review and critical reflection on equivalence and invariance tests. *Journal of Cross-Cultural Psychology*, *49*(5), 713-734.

Botcheva, L., Shih, J., & Huffman, L. C. (2009). Emphasizing cultural competence in evaluation: A process-oriented approach. *American Journal of Evaluation*, *30*(2), 176-188.

Bowen, M. L., & Tillman, A. S. (2015). Developing culturally responsive surveys: Lessons in development, implementation, and analysis from Brazil's African descent communities. *American Journal of Evaluation*, *36*(1), 25-41.

BrckaLorenz, A., Gonyea, R., & Miller, A. (2012, June). Updating the national survey of student engagement: Analyses of the NSSE 2.0 pilots. Association for Institutional Research Annual Forum.

Bresciani, M. J., Gardner, M. M., & Hickmott, J. (2009). *Demonstrating student success: A practical guide to outcomes-based assessment of learning and development in student affairs.* Sterling, VA: Stylus Publishing.

Butty, J.L., Reid, M.D., LaPoint, V. (2004). A culturally responsive evaluation approach applied to the talent development school-to-career intervention program. *New Directions for Evaluation, 101*, 37-47.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychological bulletin*, *105*(3), 456.

Byrne, B. M., Stewart, S. M., Kennard, B. D., & Lee, P. W. (2007). The Beck Depression Inventory-II: Testing for measurement equivalence and factor mean differences across Hong Kong and American adolescents. *International Journal of Testing*, *7*(3), 293-309.

Byrne, B. M., & Van de Vijver, F. J. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing*, *10*(2), 107-132.

Byrne, B.M., Van de Vijver, F.J.R. (2014). Factorial structure of the family values scale from a multilevel-multicultural perspective. *International Journal of Testing, 14*(2), 168-192.

Campbell, C. M., & Cabrera, A. F. (2011). How sound is NSSE?: Investigating the psychometric properties of NSSE at a public, research-extensive institution. *The Review of Higher Education*, *35*(1), 77-103.

Campbell, H. L., Barry, C. L., Joe, J. N., & Finney, S. J. (2008). Configural, metric, and scalar invariance of the modified achievement goal questionnaire across African

American and white university students. *Educational and Psychological Measurement*, *68*(6), 988-1007.

Canel-Çınarbaş, D., Cui, Y., & Lauridsen, E. (2011). Cross-cultural validation of the Beck depression inventory–II across US and Turkish samples. *Measurement and Evaluation in Counseling and Development*, *44*(2), 77-91.

Carrola, P. A., Yu, K., Sass, D. A., & Lee, S. M. (2012). Measurement invariance of the Counselor Burnout Inventory across cultures: A comparison of US and Korean counselors. *Measurement and Evaluation in Counseling and Development*, *45*(4), 227-244.

Carter, S. M., & Little, M. (2007). Justifying knowledge, justifying method, taking action: Epistemologies, methodologies, and methods in qualitative research. *Qualitative Health Research*, *17*(10), 1316-1328.

Cataldi, E. F., Bennett, C. T., & Chen, X. (2018). First-Generation Students: College Access, Persistence, and Postbachelor's Outcomes. Stats in Brief. NCES 2018-421. *National Center for Education Statistics*.

Cauffman, E., & MacIntosh, R. (2006). A Rasch differential item functioning analysis of the Massachusetts Youth Screening Instrument: Identifying race and gender differential item functioning among juvenile offenders. *Educational and Psychological Measurement*, *66*(3), 502-521.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural equation modeling*, *9*(2), 233-255.

Chilisa, B., & Tsheko, G. N. (2014). Mixed methods in indigenous research: Building relationships for sustainable intervention outcomes. *Journal of Mixed Methods Research*, *8*(3), 222-233.

Chouinard, J.A. (2016). Introduction: Demystifying the concept of culture in international development evaluations. *Canadian Journal of Program Evaluation*, 30.3, Special Issue, 237-247.

Chouinard, J.A., Cousins, J.B. (2009). A review and synthesis of current research on cross-cultural evaluation. *American Journal of Evaluation, 304*(4), 457-494.

Chouinard, J. A., & Cram, F. (2019). *Culturally Responsive Approaches to Evaluation* (Vol. 4). SAGE Publications.

Christie, C. A., & Alkin, M. C. (2013). An evaluation theory tree. *Evaluation roots: A wider perspective of theorists' views and influences*, 11-57. Christie and Barela (2005)

Christie, C. A., & Barela, E. (2005). The Delphi technique as a method for increasing inclusion in the evaluation process. *The Canadian Journal of Program Evaluation*, *20*(1), 105.

Clayson, Z.C., Castañeda, X., Sanchez, E., Brindis, C. (2002). Unequal power – Changing landscapes: Negotiations between evaluation stakeholders in Latino communities. *American Journal of Evaluation*, *23*(1), 33-44.

Coppens, N.M., Page, R., Chan Thou, T. (2006). Reflections on the evaluation of a Cambodian youth dance program. *American Journal of Community Psychology*, *37*(3/4), 321-331.

Corbin, J., & Strauss, A. (2014). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage publications.

Cram, F., & Mertens, D. M. (2016). Negotiating solidarity between indigenous and transformative paradigms in evaluation. *Evaluation Matters—He Take Tō Te Aromatawai*, *2*, 161-189.

Creswell, J. W. (2014). *A concise introduction to mixed methods research*. SAGE publications.

Crooks, C. V., Lapp, A., Auger, M., van der Woerd, K., Snowshoe, A., Rogers, B. J., ... & Caron, C. (2018). A Feasibility Trial of Mental Health First Aid First Nations: Acceptability, Cultural Adaptation, and Preliminary Outcomes. *American journal of community psychology*, *61*(3-4), 459-471.

Dahler-Larson, P. (2012). *The evaluation society.* Stanford; Stanford University Press.

Educational Testing Service. (2014). *ETS standards for quality and fairness*. Educational Testing Service.

Elosua, P., & López-Jaúregui, A. (2007). Potential sources of differential item functioning in the adaptation of tests. *International Journal of Testing*, *7*(1), 39-52.

Fosnacht, K., & Gonyea, R. M. (2018). The Dependability of the Updated NSSE: A Generalizability Study. *Research & Practice in Assessment*, *13*, 62-73.

Frierson, H.T., Hood, S. and Hughes, G. (2002). G.B. "Strategies that address culturally responsive evaluations." In J. Frechtling (ed.), *The 2002 User-Friendly Handbook for Project Evaluation.* Arlington, Va.: National Science Foundation, 2002.

Frierson, H., Hood, S., Hughes, G., & Thomas, V. (2010). The 2010 User-Friendly Handbook for Project Evaluation. National Science Foundation, 2010.

Garaway, G. (1996). The case-study model: An organizational strategy for cross-cultural evaluation. *Evaluation, 2*(2), 201-211.

Gibbons, M. M., Rhinehart, A., & Hardin, E. (2019). How first-generation college students adjust to college. *Journal of College Student Retention: Research, Theory & Practice, 20*(4), 488-510.

Gordon, E. W., Miller, F., & Rollock, D. (1990). Coping with communicentric bias in knowledge production in the social sciences. *Educational Researcher*, 19(3), 14-19.

Greene, J.C. (2005). Context. *Encyclopedia of Evaluation* (pp. 83-85). Thousand Oaks. SAGE Publications, Inc.

Greene, J. C. (2007). *Mixed methods in social inquiry* (Vol. 9). John Wiley & Sons.

Greene, J. C. (2015). The emergence of mixing methods in the field of evaluation. *Qualitative Health Research, 25*(6), 746-750.

Grover, R. K., & Ercikan, K. (2017). For which boys and which girls are reading assessment items biased against? Detection of differential item functioning in heterogeneous gender populations. *Applied Measurement in Education, 30*(3), 178-195.

Guttmannova, K., Szanyi, J.M., Cali, P.W. (2008). Internalizing and externalizing behavior problem scores: Cross-ethnic and longitudinal measurement invariance

of the behavior problem index. *Educational and Psychological Measurement, 68*(4), 676-694.

Guzmán, B.L. (2003). Examining the role of cultural competency in program evaluation: Visions for new millennium evaluators. In S.I. Donaldson & M. Scriven (Eds.), *Evaluating social programs and problems: Visions for the new millennium* (pp. 167-181). Mahwah, NJ: Lawrence Erlbaum Associates.

Haeger, H., Lambert, A. D., Kinzie, J., & Gieser, J. (2012, June). Using cognitive interviews to improve survey instruments. Association for Institutional Research Annual Forum.

Hall, S. (1992). The west and the rest: Discourse and power. In S. Hall & B. Gieben (Eds.), *Formations of modernity* (pp. 275-331). Cambridge, GB: Polity Press.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Eds.). (2004). *Adapting educational and psychological tests for cross-cultural assessment*. Psychology Press.

He W., Wolfe, E.W. (2010). Item equivalence in English and Chinese translation of a cognitive development test for preschoolers. *International Journal of Testing, 10*(1), 80-94.

Hopson, R. K. (2009). Reclaiming knowledge at the margins: Culturally responsive evaluation in the current evaluation moment. *The Sage international handbook of educational evaluation*, 429-446.

Hood, S. (2004). A journey to understand the role of culture in program evaluation: Snapshots and personal reflections of one African American evaluator. *New Directions for Evaluation*, *2004*(102), 21-37.

Hood, S., Frierson, H., & Hopson, R. (Eds.). (2005). *The role of culture and cultural context in evaluation: A mandate for inclusion, the discovery of truth and understanding*. IAP.

Hood, S., Hopson, R., & Frierson, H. (2015). Introduction: This is where we continue to stand. *Continuing the journey to reposition culture and cultural context in evaluation theory and practice*.

Hood, S., Hopson, R. K. and Kirkhart, K. E. (2015) Culturally Responsive Evaluation, in *Handbook of Practical Program Evaluation*, Fourth (eds K. E. Newcomer, H. P. Hatry and J. S. Wholey), John Wiley & Sons, Inc., Hoboken, NJ, USA.

Huang, X., Wilson, M., & Wang, L. (2016). Exploring plausible causes of differential item functioning in the PISA science assessment: language, curriculum or culture. *Educational Psychology*, *36*(2), 378-390.

Hughes, D., Seidman, E., Williams, N. (1993). Cultural phenomena and the research enterprise: Toward a culturally anchored methodology. *American Journal of Community Psychology, 21*(6), 687-703.

Janzen, R., Ochocka, J., & Stobbe, A. (2016). Towards a theory of change for community-based research projects. *Engaged Scholar Journal: Community-Engaged Research, Teaching, and Learning*, *2*(2), 44-64.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2020). semTools: Useful tools for structural equation modeling. R package version 0.5-3. Retrieved from https://CRAN.R-project.org/package=semTools

Kane, M. T. (2013). Validation as a pragmatic, scientific activity. *Journal of Educational Measurement*, *50*(1), 115-122.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

Kato, K., Moen, R. E., & Thurlow, M. L. (2009). Differentials of a state reading assessment: Item functioning, distractor functioning, and omission frequency for disability categories. *Educational Measurement: Issues and Practice, 28*(2), 28-40.

Kirkhart, K. (1995). Seeking multicultural validity: A postcard from the road. *Evaluation Practice, 16,* 1-12.

Kirkhart, K. E. (2005). Through a cultural lens. *The Role of Culture and Cultural Context: A Mandate for Inclusion, the Discovery of Truth and Understanding in Evaluative Theory and Practice,* 21.

Kirkhart, K. (2010). Eyes on the prize: Multicultural validity and evaluation theory *American Journal of Evaluation, 31*(3), 400-413.

Kirkhart, K. E. (2013, April). Repositioning validity. In *Plenary on Perspectives on Repositioning Culture in Evaluation and Assessment at CREA Inaugural Conference,* Chicago.

Kline, R. B. (2011). *Principles and practice of structural equation modeling. (3nd).* The Guildford Press.

Kornilov, S. A., Kornilova, T. V., & Grigorenko, E. L. (2016). The Cross-Cultural Invariance of Creative Cognition: A Case Study of Creative Writing in US and Russian College Students. *New directions for child and adolescent development*, *2016*(151), 47-59.

Kuh, G. D. (2001). Assessing what really matters to student learning inside the national survey of student engagement. *Change: The Magazine of Higher Learning*, *33*(3), 10-17.

Kuh, G. D. (2008). High-impact educational practices: What they are, who has access to them, and why they matter. Washington, DC: Association of American Colleges and Universities.

Kuh, G.D. (2009). The national survey of student engagement: Conceptual and empirical foundations. *New Directions for Institutional Research,* 141, 5-20.

Kuh, G. D., Cruce, T. M., Shoup, R., Kinzie, J., & Gonyea, R. M. (2008). Unmasking the effects of student engagement on first-year college grades and persistence. *The journal of higher education*, *79*(5), 540-563.

Kuh, G. D., Kinzie, J., Cruce, T., Shoup, R., & Gonyea, R. M. (2007). *Connecting the dots: Multi-faceted analyses of the relationships between student engagement results from the NSSE, and the institutional practices and conditions that foster student success*. Indiana University Center for Postsecondary Research.

Kuh, G. D., Jankowski, N., Ikenberry, S. O., & Kinzie, J. (2014). *Knowing what students know and can do: The current state of student learning outcomes assessment in U.S. colleges and universities.* Champaign, IL: National Institute for Learning Outcomes Assessment.

Lakin, J. M., Elliott, D. C., & Liu, O. L. (2012). Investigating ESL students' performance on outcomes assessments in higher education. *Educational and Psychological Measurement, 72*(5), 734-753.

LaNasa, S. M., Cabrera, A. F., & Trangsrud, H. (2009). The construct validity of student engagement: A confirmatory factor analysis approach. *Research in Higher Education, 50*(4), 315-332.

LaPoint, V., & Jackson, H. L. (2004). Evaluating the co-construction of the family, school, and community partnership program in a low-income urban high school. *New Directions for Evaluation, 2004*(101), 25-36.

Lau, A. L., Cummins, R. A., & Mcpherson, W. (2005). An investigation into the cross-cultural equivalence of the Personal Wellbeing Index. *Social Indicators Research, 72*(3), 403-430.

Li, C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior research methods, 48*(3), 936-949.

Lincoln, Y.S., Guba, G.G. (1985). *Naturalistic Inquiry.* Newburry Park, London: Sage

Liu, O.L., Wilson, M. (2009). Gender differences and similarities in PISA 2003 mathematics: A comparison between the United States and Hong Kong. *International Journal of Testing*, *9*(1), 20-40.

Luyt, R. (2012). A framework for mixing methods in quantitative measurement development, validation, and revision: A case study. *Journal of Mixed Methods Research*, *6*(4), 294-316.

Maddox, B., Zumbo, B., Tay-Lim, B., Qu, D. (2015). An anthropologist among the psychometricians: Assessment events, ethnography, and differential item functioning in the Mongolian Gobi. *International Journal of Testing*, *15*(4), 291-309.

Maki, P. L. (2012). *Assessing for learning: Building a sustainable commitment across the institution*. Stylus Publishing, LLC.

Mamaril, M. N., Cox, L. J., & Vaughan, M. (2018). Weaving evaluation into the Waipā ecosystem: Placing evaluation in an indigenous place-based educational program. *Studies in Educational Evaluation*, *56*, 42-51.

Martinez, A., Running Wolf, P., BigFoot, D. S., Randall, C., & Villegas, M. (2018). The process of becoming: A roadmap to evaluation in Indian country. *New Directions for Evaluation*, *2018*(159), 33-45.Maxwell, J.A. (2013). *Qualitative research design: An interactive approach.* Los Angeles, CA: Sage.

McFarland, J., Hussar, B., Wang, X., Zhang, J., Wang, K., Rathbun, A., & Mann, F. B. (2018). The Condition of Education 2018. NCES 2018-144. *National Center for Education Statistics*.

Means, D. R., & Pyne, K. B. (2017). Finding my way: Perceptions of institutional support and belonging in low-income, first-generation, first-year college students. *Journal of College Student Development*, *58*(6), 907-924.

Meredith, W. (1993). Measurement invariance, factor analysis and invariance. *Psychometrika*, *58*(4), 525-543.

Mertens, D.M. (2007). Transformative paradigm: Mixed methods and social justice. *Journal of Mixed Methods Research*, 1(3), 212-225.

Mertens, D. M. (2010). Transformative mixed methods research. *Qualitative inquiry*, *16*(6), 469-474.

Mertens, D. M. (2011). Mixed Methods as Tools for Social Change. *Journal of Mixed Methods Research*, *5*(3), 195–197.

Mertens, D.M., & Wilson, A. T. (2012). *Program Evaluation Theory and practice* (Chapter six). New York, NY: The Guilford Press.

Mertens, D. M., & Wilson, A. T. (2018). *Program evaluation theory and practice*. Guilford Publications.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational measurement: Issues and practice*, *14*(4), 5-8.

Miller, A. L., Sarraf, S. A., Dumford, A. D., & Rocconi, L. M. (2016). Construct validity of NSSE engagement indicators (NSSE psychometric portfolio report). Bloomington, IN: Center for Postsecondary Research, Indiana University, School of Education.

Mohatt, G. V., Hazel, K. L., Allen, J., Stachelrodt, M., Hensel, C., & Fath, R. (2004). Unheard Alaska: Culturally anchored participatory action research on sobriety with Alaska Natives. *American Journal of Community Psychology*, *33*(3-4), 263-273.

Montenegro, E., & Jankowski, N. A. (2017). Equity and Assessment: Moving towards Culturally Responsive Assessment. Occasional Paper# 29. *National Institute for Learning Outcomes Assessment*.

Mylonas, K., Furnham, A., & Country Collaborators. (2014). Bias in terms of culture and a method for reducing it: An eight-country "Explanations of Unemployment Scale" study. *Educational and Psychological Measurement*, *74*(1), 77-96.

Nastasi, B. K., & Hitchcock, J. H. (2016). *Mixed methods research and culture-specific interventions: Program design and evaluation* (Vol. 2). SAGE Publications.

National Center for Education Statistics. (2016). Table 326.10. Graduation rate from first institution attended for first-time, full-time bachelor's degree- seeking students at 4-year postsecondary institutions, by race/ethnicity, time to completion, sex, control of institution, and acceptance rate: Selected cohort entry years, 1996 through 2009. Retrieved from: https://nces.ed.gov/programs/digest/d16/tables/dt16_326.10.asp

National Center for Education Statistics. (2019). Profile of Undergraduate Students: Attendance, distance and remedial education, degree program and field of study, demographics, financial aid, financial literacy, employment and military status:

2015-2016. Retrieved from: https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2019467

National Survey of Student Engagement. (2018). NSSE Conceptual Framework (2018) (NSSE Psychometric Portfolio Report). Bloomington, IN: Center for Postsecondary Research, Indiana University, School of Education. Retrieved from: http://nsse.indiana.edu/html/psychometric_portfolio.cfm

NSSE - National Survey of Student Engagement. (n.d.). Retrieved August 02, 2019, from: http://nsse.indiana.edu/html/about.cfm

Oliveri, M. E., Ercikan, K., Lyons-Thomas, J., & Holtzman, S. (2016). Analyzing fairness among linguistic minority populations using a latent class differential item functioning approach. *Applied Measurement in Education*, *29*(1), 17-29.

Ouimet, J.A., Bunnage, J.C., Carini, R.M., Kuh, G.D., Kennedy, J. (2004). Using focus groups, expert advice, and genitive interviews to establish the validity of a college student survey. *Research in Higher Education, 45*(3), 233-250.

Pacico, J.C., Zanon. C, Bastianello, M., Reppold, C.T., Hutz, C. (2013) Adaptation and Validation of the Brazilian Version of the Hope Index, *International Journal of Testing, 13*(3), 193-200.

Pascarella, E. T., Pierson, C. T., Wolniak, G. C., & Terenzini, P. T. (2004). First-generation college students: Additional evidence on college experiences and outcomes. *The Journal of Higher Education*, *75*(3), 249-284.

Patton, M.Q. (2008). *Utilization-focused evaluation* (4th ed). Chapter 2: What is utilization-focused evaluation? How do you get started? Thousand Oaks, CA: Jossey-Bass.

Penfield, R. D. (2016). Fairness in test scoring. In L. L. Cook & N. J. Dorans (Eds.), *Fairness in educational assessment and measurement* (pp. 55–75). New York, NY: Routledge.

Pike, G. R., & Kuh, G. D. (2005). First-and second-generation college students: A comparison of their engagement and intellectual development. *The Journal of Higher Education*, *76*(3), 276-300.

Pokropek, A., Borgonovi, F., & McCormick, C. (2017) On the Cross-Country Comparability of Indicators of Socioeconomic Resources in PISA, *Applied Measurement in Education, 30*(4), 243-258.

Porter, S.R. (2011). Do college student surveys have any validity? *The Review of Higher Education*, *35*(1), 45-76.

Prelow, H. M., Tein, J. Y., Roosa, M. W., & Wood, J. (2000). Do coping styles differ across sociocultural groups? The role of measurement equivalence in making this judgment. *American Journal of Community Psychology*, *28*(2), 225-244.

Psaki, S. R., & Hindin, M. J. (2016). Lessons in cross-cultural measurement of depressive symptoms: findings from a mixed-methods study in Ghana. *International journal of culture and mental health*, *9*(4), 340-355.

Randall, J., Sireci, S., Li, X., & Kaira, L. (2012). Evaluating the Comparability of Paper- and Computer-Based Science Tests Across Sex and SES Subgroups. *Educational Measurement: Issues and Practice*, *31*(4), 2-12.

Redford, J., & Mulvaney Hoyer, K. (2017). First Generation and Continuing-Generation College Students: A Comparison of High School and Postsecondary Experiences.

Ross, S.J., Okabe, J. (2006). The subjective and objective interface of bias detection on language tests. International Journal of Testing, 6(3), 229-253.

Rosseel Y (2012). "lavaan: An R Package for Structural Equation Modeling." *Journal of Statistical Software*, 48(2), 1–36. Retrieved from: http://www.jstatsoft.org/v48/i02/.

Scriven, M. (1991). *Evaluation thesaurus*. Sage.

SenGupta, S., Hopson, R., Thompson-Robinson, M. (2004). Cultural competence in evaluation: An overview. *New Directions for Evaluation*, 102, 5-19.

Small, S.A., Tiwari, G., Huser, M. (2006). The cultural education of academic evaluators: Lessons from a University-Hmong community partnership. *American Journal on Community Psychology, 37*, 357-364.

Stanfield, J.H. Jr. (1999). Slipping through the front door: Relevant social scientific evaluation in the people of color century. *American Journal of Evaluation* 20(3), 415-431.

Stokes, H., Chaplin, S.S., Dessouky, S., Aklilu, L., Hopson, R.K. (2011). Addressing social injustices, displacement, and minority rights through cases of culturally

responsive evaluation. *Diaspora, Indigenous, and Minority Education, 5,* 167-177.

Sutton, S., Baxter, N., Grey, K., & Putt, J. (2016). Working both-ways: Using participatory and standardised methodologies with Indigenous Australians in a study of remote community safety and wellbeing. *Evaluation Journal of Australasia, 16*(4), 30-40.

Sy, A., Greaney, C, Nigg, Hirose-Wong, S.M. (2015). Developing a measure to evaluate a positive youth development program for Native Hawaiians: The Hui Malama o ke Kai rubrics of Hawaiian values. *Asia-Pacific Journal of Public Health, 27*(2), NP1517-NP1528.

Symonette, H. (2004). Walking pathways toward becoming a culturally competent evaluator: Boundaries, borderlands, and border crossings. *New Directions for Evaluation, 2004*(102), 95-109.

Thelin, J. R. (2019). *A history of American higher education*. (3rd ed.). Baltimore, MD: Johns Hopkins University Press.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of educational research, 45*(1), 89-125.

Uhl, G., Robinson, B., Westover, B., Bockting, W., & Cherry-Porter, T. (2004). Involving the community in HIV prevention program evaluation. *Health promotion practice, 5*(3), 289-296.

Unger, J. B., Soto, C., & Thomas, N. (2008). Translation of health programs for American Indians in the United States. *Evaluation & the health professions*, *31*(2), 124-144.

US News Education | Best Colleges | Best Graduate Schools | Online Schools. (2019). Retrieved February 02, 2019, from https://www.usnews.com/education?int=top_nav_Education

Vogel, D. L., Armstrong, P. I., Tsai, P. C., Wade, N. G., Hammer, J. H., Efstathiou, G., ... & Topkaya, N. (2013). Cross-cultural validity of the Self-Stigma of Seeking Help (SSOSH) scale: Examination across six nations. *Journal of counseling psychology*, *60*(2), 303.

Ward, L., Siegel, M. J., & Davenport, Z. (2012). *First-generation college students: Understanding and improving the experience from recruitment to commencement*. John Wiley & Sons.

Whitley, S. E., Benson, G., & Wesaw, A. (2018). First-generation student success: A landscape analysis of programs and services at four-year institutions. *NASPA—Student Affairs Administrators in Higher Education. Retrieved from https://firstgen.naspa.org/2018- landscape-analysis*.

Wu, P. C. (2010). Measurement invariance and latent mean differences of the Beck Depression Inventory II across gender groups. *Journal of Psychoeducational Assessment*, *28*(6), 551-563.

Zerquera, D., Reyes, K. A., Pender, J. T., & Abbady, R. (2018). Understanding practitioner- driven assessment and evaluation efforts for social justice. *New Directions for Institutional Research, 2018*(177), 15-40.

Zilvinskis, J., Masseria, A. A., & Pike, G. R. (2017). Student engagement and student learning: Examining the convergent and discriminant validity of the revised national survey of student engagement. *Research in Higher Education, 58*(8), 880-903.

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF). *Ottawa: National Defense Headquarters*.

APPENDIX A. CRITICAL EXAMINATION OF THE LITERATURE

Table 12: Matrix from Critical Examination of the Literature

Article	**Program/Context Description**	**Population**	**Data collection**	**Strategies for cultural relevance or invariance**	**Analysis**
Abubakar & van de Vijver (2017)	This article focused on test adaptation, adoption, and assembly, examining equivalence in adopted or adapted tests, and recommended a four-stage model for systematic assembly or adaptation in cross-cultural contexts including examinations of equivalence.	Sub-Saharan Africa	Survey	Recommended steps from the four-stage model: Review the literature for existing instruments and make cultural modifications, collect qualitative data through community involvement about constructs, conduct panel reviews for quality of translations, use cognitive interviews to examine quality of instrument, test the psychometric (including equivalence) and non-psychometric properties	CFA

Article	Program/Context Description	Population	Data collection	Strategies for cultural relevance or invariance	Analysis
Alkon et al. (2001)	This study used multiple methods to explore a violence-prevention program serving ethnically diverse children and families in a child care environment.	English, Chinese, and Spanish speaking families	Surveys, focus groups, observations, interviews	Used the following criteria to examine instruments prior to use: language at a fifth-grade reading level, conceptual relevance to the research questions, cultural relevance; sought to establish validity of data collected across three diverse ethnic populations, emphasis on establishing conceptual equivalence	Analysis not discussed
Allen et al (2014)	Allen et al. (2014), empirically tested a culturally grounded intervention program designed to support well-being, defined as reasons for life and sobriety, with Alaska Native youth using culturally sensitive and appropriate measures.	Alaska Native youth	Survey	Locally defined construct with community engagement, involved Yup'ik cultural consultants, cultural review of items and pilot testing of revised, adopted measure. Used a strengths-based approach and positive psychology	CFA to determine dimensionality, IRT to identify best functioning items

Article	Program/Context Description	Population	Data collection	Strategies for cultural relevance or invariance	Analysis
Asil & Brown (2016)	Authors examined equivalence in the PISA Reading Comprehension test by language, culture, and economic development across 55 countries.	Multiple countries using Australia as the baseline	Survey	Framed their equivalence study using Hofstede's (2007) cultural attributes theory and included discussions of culture in their findings	MGCFA
Ausili et al. (2019)	The purpose of this study was to test the generalizability and comparability, through testing the invariance, of the Self-Care of Diabetes Inventory across cultures and languages, specifically between Italy and the United States.		Survey	Examined equivalence, discussed the sociocultural differences between the United States and Italy	MGCFA
Banks (2006)	Presents a working definition of the term culture, a framework for evaluating cultural bias in educational testing, and applies the framework.	Hispanic, Black, and White students	Survey	Used cultural taxonomies as part of a larger framework to evaluate hypothesis testing about cultural bias	DBF, DDF
Banks (2012)	Presents a 7 step process for identifying if inferential reading items are more prone to cultural bias than literal reading items across Hispanic, Black, and White students in the Midwest	Hispanic, Black, and White students	Survey	Used the literature to identify cultural aspects that describe group cultures as a framework for the analysis	DBF, DIF, DDF

Article	Program/Context Description	Population	Data collection	Strategies for cultural relevance or invariance	Analysis
Bodkin-Andrews et al. (2010)	Examined cross-cultural equivalence through invariance testing & latent mean differences for questionnaire with Indigenous and non-Indigenous Australian secondary school students.	Indigenous and non-Indigenous Australian students, males and females	Survey	Examined invariance across multiple groups	CFA for factor structure, invariance testing, multiple indicator, multiple causes modeling to test for differences across gender, ethnicity, and the interaction of these two variables
Botcheva et al. (2009)	This article focuses on culturally competent evaluation when evaluating an HIV/AIDS program in Zimbabwe including the revision of an existing survey to make it more culturally appropriate.	Students in Zimbabwe	Survey, poems	Reviewed existing literature, consulted with program providers, integrated cultural characteristics into the survey, changed response preferences from multiple choice to true/false, removed culturally irrelevant concepts, adapted survey to reflect culture using vignettes	content analysis for poems

Article	Program/Context Description	Population	Data collection	Strategies for cultural relevance or invariance	Analysis
Bowman & Tillman (2015)	The development and implementation of three culturally responsive surveys in an international context.	*Quilombos* in Brazil	Surveys	Focus groups, adopted and adapted an instrument used in international contexts, stakeholders were critical in providing feedback on the surveys	Data entered into a data base for constructs to be analyzed
Butty et al. (2004)	This article focuses on the successes and challenges encountered in evaluating Howard University's Research on the Education of Students Placed At Risk (CRESPAR) urban school to career intervention program using culturally responsive evaluation.	African American 9th graders, urban environment	Surveys, interviews, focus groups, pre and posttest, self-assessments, standardized assessments	Instruments were selected after reviewing them for cultural sensitivity in form, language, and content, used multiple methods, disaggregated data, engaged stakeholder in data analysis and interpretation	Data was disaggregated
Byrne & van de Vijver (2010)	Researchers examined the Family Values Scale across 30 countries using SEM approaches to test for multi-group equivalence, identified problems with this approach, and recommended a multipronged approach to examining equivalence focusing on country and scale items as possible sources of bias.	Culture is a function of country in this study	Survey	Examined equivalence across 30 countries	MGCFA

Article	Program/Context Description	Population	Data collection	Strategies for cultural relevance or invariance	Analysis
Byrne & van de Vijver (2014)	Focus was to validate the factor structure of the Family Values Scale, test equivalence across countries, and add covariates of gender at the individual level and religion and affluence at the country level.	Culture is a function of country in this study	Survey	Examined equivalence across 27 countries, adding individual characteristics such as gender	MGCFA
Byrne et al. (2007)	The purpose of this study was to examine equivalence in the Beck Depression Inventory-II in order to examine factor mean differences for Hong Kong and American adolescents for the factors of negative attitude, performance difficulty, somatic elements, and general depression.	Chinese and American adolescents	Survey	Examined for equivalence, contextualized findings of significant mean differences within the cultures of the participants	
Canel-Cinaras et al. (2011)	This study examined the Beck Depression Inventory-II for invariance using a samples of college students from Turkey and the US, and upon finding lack of scalar invariance, conducted differential item functioning.	U.S. and Turkish college students	Survey	Examined equivalence, differential item functioning, then worked with an expert to investigate causes of DIF	MGCFA, DIF

Article	Program/Context Description	Population	Data collection	Strategies for cultural relevance or invariance	Analysis
Carrola et al. (2012)	Examined the Counselor Burnout Inventory for measurement invariance across U.S. and Korean counselors and found that three of the five factors measured by the instrument were invariant across cultures.	White, non-Hispanic, U.S. and Korean counselors	Survey	Examined for equivalence, contextualized the findings in cultural understanding	MGCFA
Cauffman & MacIntosh (2006)	This study used a Rasch analysis to identify differential item functioning across race and gender in juvenile offenders using the Massachusetts Youth Screening Instrument, a mental health assessment.	Adolescents in the juvenile system across race and gender	Survey	Examined DIF and grounded the findings in cultural understanding	DIF
Christie & Barela (2005)	In this article, the Delphi Technique is described as a way to enhance marginalized group participation in an evaluation. This approach allows for all stakeholders to have a voice and historically marginalized voices to have equal influence.	Historically marginalized populations in college access programs	Survey and interviews	Evaluators leverage surveys for feedback to reduce power dynamics and equalize which voices are heard, stakeholders are involved in determining survey structure, content, and scale	Analysis strategies not discussed
Clayson et al. (2009)	This article discusses the multiple roles evaluators play in navigating contexts between funders and cultural community stakeholders in evaluations.	Low-income Latino community members in California	Surveys, interviews, focus groups, observations, photography	Cross-cultural evaluation team	Analysis strategies not discussed

Article	Program/Context Description	Population	Data collection	Strategies for cultural relevance or invariance	Analysis
Cokely (2014)	This study examined the academic motivations of African American college students in the context of reducing the higher education achievement gap.	African American college students	Survey	Examined factor structure	CFA
Conner (2004)	This is a case study of the culturally responsive evaluation of two programs designed to serve Latino populations in the United States with the focus of HIV prevention and incorporating aspects of multicultural validity.	Latino immigrant farmworkers	Icon matching surveys, pretest posttest	Piloted, significant input from migrant workers, discussed multicultural validity	Analysis strategies not discussed
Coppens et al. (2006)	This article focuses on the culturally responsive evaluation of a Cambodian dance program north of Boston in the United States and the dynamic of using a funder mandated standardized survey with a more culturally responsive and locally developed evaluation approach.	Cambodian youth	Interviews, survey, focus groups, survey	Administered the required survey as well as a locally developed survey	Analysis strategies not discussed
Crooks et al. (2018)	This is a mixed methods approach to evaluating health outcomes in First Nations.	First Nations people in Canada	Surveys, interviews	7 out of 8 evaluators were Indigenous, prioritized Indigenous knowledge to guide work	ANOVA, t-test, qualitative coding

Article	Program/Context Description	Population	Data collection	Strategies for cultural relevance or invariance	Analysis
Demir (2016)	The purpose of this study was to build a multivariate structure which modeled school characteristics and then to test the invariance of this structure across five countries using PISA 2012 science data; the model was constructed using the PISA school questionnaire.	Culture is a function of country in this study	Survey	Discussed how culture influences education	MGCFA
Elosua & Lopez-Jauregui (2007)	The study aimed to classify sources of differential item functioning resulting from an aptitude test given in Spanish and Basque populations using two separate panels for judgmental reviews and DIF analysis for statistical review.	Culture focused on Spanish and Basque respondents	Survey	Two panels of experts reviewed the items, DIF analysis conducted, cultural relevance in the items was a central part of the discussion regarding test adaptation across Basque and Spanish samples	Judgmental reviews and statistical analysis
He & Wolfe (2010)	Authors in this study examined nonequivalence between English and Chinese versions of a cognitive assessment for children. Both judgmental and statistical methods for examining item bias.	American and Chinese preschoolers	Survey	Researchers from each country helped design, test, and revise the instrument. Test was administered verbally in the children's home language. Examined if items were overly familiar to children based on socio-cultural experiences.	Judgmental reviews and statistical analysis

Article	Program/Context Description	Population	Data collection	Strategies for cultural relevance or invariance	Analysis
Hjemdal et al. (2011)	Authors explored the construct validity of the Resilience Scale for Adults by conducting a series of invariance tests across Belgian and Norwegian samples. Evidence supported metric invariance for five out of six factors measured.	Belgian and Norwegian respondents	Survey	Talked about cultural differences in the two populations	MGCFA
Huang et al. (2016)	Examined the PISA science assessment for differential item functioning across language, curriculum, and culture.	Samples from the U.S., Canadian, Chinese Hong Kong, and mainland Chinese students	Survey	Examined instrument for differential functioning across countries, examined cultural differences as a potential cause of DIF, panelist examined items, panelist included 15 year olds (the target audience for the test), panelist included content experts	DIF analysis

Article	Program/Context Description	Population	Data collection	Strategies for cultural relevance or invariance	Analysis
Jenzen et al. (2015)	This study started with a survey instrument to measure the constructs of hope and resilience in a faith based intervention program for youth. The scale was selected and feedback was solicited from the steering committee and students. A series of validity and reliability tests were conducted and findings determined that the hope scale determined poor fit. Using qualitative data from the population, a new set of items was constructed which determined much better fit.	Inner-city youth, low-income areas of Hamilton Ontario	Survey	Evaluator and stakeholders agreed on the survey design, there was a steering committee that agreed on the questions, pilot tested the revised instrument, recognized the theory and tool were developed and validated using a different demographic than the program was designed to serve, hope on the survey was too narrowly defined for this population	EFA, CFA, stakeholder feedback, pilot interviews
Kornilov et al. (2016)	Investigated invariance of creativity measure across Russian Federation and U.S. students.	Russian Federation and U.S. students	Creative writing story scores given by experts	Provided a discussion of how culture shapes creativity, examined construct invariance, positioned their findings in a discussion of cultural attributes	DIF, CFA

Article	Program/Context Description	Population	Data collection	Strategies for cultural relevance or invariance	Analysis
Koss et al., (2003)	Examined impact of childhood exposures on alcohol dependence.	Native American participants across 7 tribes	Fact to face interviews (1,660) then logistic regression	Included multiple tribes, examined data within and across tribes, discussed needing a more nuanced understanding of tribal characteristics to improve their study (e.g., tribal integration, degree of ancestry, lifestyle characteristics)	Logistic regression
Lakin et al. (2012)	This study examined the validity of a measure of higher education outcomes for English as Second Language students.	English as Second Language Students	Survey	Conducting invariance testing, questioned the validity of a standardized tool for ESL students	MGCFA, DIF, DBF
LaPoint & Jackson (2004)	Evaluate a program impact on student academic achievement and social impact.	Black, low-income, urban high-school students	Surveys, focus groups, interviews	Stakeholder involvement including students, evaluators shared some of the identities of the students, stakeholder reviewed and refined questions for assessments, piloted measures	Analysis strategies not discussed

Article	Program/Context Description	Population	Data collection	Strategies for cultural relevance or invariance	Analysis
Lau (2005)	Examined the Personal Wellbeing Index (PWI) for cross-cultural invariance across samples of Hong Kong Chinese and Australian samples. Difference were found and attributed to cultural response bias.	Hong Kong Chinese and Australian participants	Survey via telephone interview	Questioned validity of tool across cultural populations, discussed cultural differences in the conceptualization of constructs on the scale (e.g., well-being)	EFA, MANOVA
Luyt (2012)	Used a case study of an evaluation of Male Attitudes in different cultures to showcase the role of multiple methods in cross-cultural assessment.	Afrikaans, English, and Xhosa versions of the measure	Survey, focus groups, factor analysis	Qualitative data used to inform if the use of the tool across cultures was appropriate, quantitative data used to determine construct validity, culture was considered as part of the reason for differences in the data	Replicatory factor analysis
Maddox (2015)	This article combines ethnography with differential item function in order to develop a more holistic and comprehensive understanding of reading literacy in the Mongolian Gobi as a part of the UNESCO efforts.	Mongolian herders	Ethnography, survey	Engaged ethnography experts to shape understandings of the performance of the test in a given context, examined how respondents coped with items with varying degrees of relevance to herders cultural context	Qualitative analysis, DIF

Article	Program/Context Description	Population	Data collection	Strategies for cultural relevance or invariance	Analysis
Mamaril Cox Vaughan (2018)	The authors of this article examined how a logic model, a traditionally Western tool, could be used to evaluate a Native Hawaiian education program.	Native Hawaiian families	Parent survey, participant talk story sessions, interviews	Included methods which were culturally congruent	Analysis strategies not discussed
Marsh et al. (2006)	Authors examined the Student Approaches to Learning instrument for invariance across 25 countries and included additional criterion variables of gender, socioeconomic status, math achievement, and verbal achievement.	25 countries and included gender, socioeconomic status	Survey	Examined construct validity and cross-cultural generalizability, conducted invariance testing	CFA
Mertens & Hopson (2006)	This article provides and overview of a number of studies in S.T.E.M fields which support diverse student learners and used a culturally responsive evaluation approach.	Reflection of multiple programs working with diverse youth and Native American students	Multiple tools used across different programs	Questions the use of standardized instruments in highly diverse communities, emphasized importance of cultural competence, importance of implications of evaluation work in a context	Analysis strategies not discussed
Miller & Lee (2009)	This study examines the invariance of the Asian American Family Conflicts Scale for different Asian American ethnic groups.	Asian Americans and Chinese National samples	Survey	Examines survey for factorial invariance (measurement and structural), discusses within group analysis	MGCFA

Article	Program/Context Description	Population	Data collection	Strategies for cultural relevance or invariance	Analysis
Mohatt et al. (2004)	Used culturally anchored participatory action research to guide the evaluation of an alcohol prevention and treatment program for Alaska Natives. They discussed the tensions and resolutions of grounding their methodology in the culture and community in which the evaluation was occurring.	Alaskan Native participants	Survey and life history methodology	Strengths based approach, involved a cultural council, chose methods which honored cultural traditions, piloted instrument, used a culturally grounded an indigenous framework to guide analysis	Coded life history data with stakeholders
Osterland et al. (2009)	The focus of this article was on translating MBTI to Chinese while maintaining construct validity.	Chinese male respondents	Survey	Examined item types across cultures and sought to confirm the factor structure	Factor analysis
Prelow et al. (2000)	This article focused on determining measurement equivalence for a measure of coping skills across Mexican American/Mexican immigrant mothers and Anglo middle-class divorced mothers using multi-group confirmatory factor analysis.	Mexican American and Mexican immigrant mothers	Survey	Questioned the validity of a standardized measure within a cultural population, discussed cultural characteristics which may influence a measure of hope/coping skills, recognized lack of reliability/validity information for diverse populations, linked findings back to discussion of culture	CFA

Article	Program/Context Description	Population	Data collection	Strategies for cultural relevance or invariance	Analysis
Psaki et al. (2016)	Used mixed methods to evaluate a mental health scale used in Kumasi, Ghana.	participants from Ghana	Focus groups, survey	Questioned the validity of a standardized measure within a cultural population, used multiple methods to investigate cultural understanding of tool	CFA
Ross & Okabe (2006)	In this study, the authors compared the bias detection techniques of panel reviewers and statistical approaches for detecting differential item functioning in a language test for Japanese university students.	Japan, Higher Education, Language testing	Survey	Conducted a judgmental review with a panel of Japanese teachers, found that cultural bias of judgmental panelist may have influenced determinations of why items were biased	DIF, judgmental review
Segeritz & Pant (2013)	This study examined the equivalence of the PISA student approaches to learning measure across cultural groups within a country. Significant variation in measurement invariance was found in one or more scales across groups examined.	German, and Turkish and USSR immigrants in Germany	Survey	Questioned the invariance of survey across cultural groups, discussed how Western cultures may conceptualize constructs differently	MGCFA

Article	Program/Context Description	Population	Data collection	Strategies for cultural relevance or invariance	Analysis
Small et al. (2006)	Reflective article on evaluating a family strengthening program for Hmong families.	Hmong family participants	Interviews, field notes, observation	Administered questionnaire as interview to navigate concerns related to cultural responsiveness, culturally foreign content of items, tool did not meet criteria for multicultural validity	Analysis strategies not discussed
Stokes et al. (2011)	Case study analysis of two interns beginning their journeys as CRE practitioner. One worked with survivors of torture and the other intern worked for the international rescue committee working with refugees.	Torture survivors and refugees	Surveys, focus groups, observations, interviews	Use of multiple methods with an understanding of the limitations of surveys to fully capture context	Analysis strategies not discussed
Sun & Hernandez (2015)	The achievement goal questionnaire demonstrates configural, metric, scalar and structural equivalence using confirmatory factor analysis across a sample of American, Chinese, and Dutch college students.	American, Chinese and Dutch college students	Survey	Examined invariance, discussed individualistic and collectivist cultural aspects	MGCFA

Article	Program/Context Description	Population	Data collection	Strategies for cultural relevance or invariance	Analysis
Sy et al. (2015)	This article focuses on the development of rubrics and 8 other tools to measure youth knowledge of indigenous values for a culturally relevant and positive youth program in a Native Hawaiian community, validity and reliability of the rubrics and other tools are discussed.	Native Hawaiian community members	Rubric, survey, interview, journals	Aimed to design a culturally relevant tool, modified through community input, multiple methods to validate measure, discussed acculturation stress, "local wisdom" approach with quantitative and qualitative data sources, include cultural factors in the measures	reliability, interclass correlation, univariate and bivariate statistical procedures, thematic coding
Unger et al. (2008)	This article focuses on the evaluation of the unique health and social challenges facing American Indian and Alaska Native teenagers and makes recommendations about conducting such evaluations including using culturally appropriate measures.	American Indian and Alaska Native teenagers	Survey, talking circle, focus groups	Included a long historical overview of American Indian and Alaska Native history, provided rich discussion of cultural characteristics, importance of culturally resonant approaches	Analysis strategies not discussed
Vogel et al. (2013)	The purpose of this article was to examine the validity, specifically measurement invariance, of a self-help stigma scale across multiple countries.	Six countries	Survey	Questioned the invariance of a survey used across cultures, discussion of findings related back to culture	MGCFA

Article	Program/Context Description	Population	Data collection	Strategies for cultural relevance or invariance	Analysis
Wang et al. (2012)	The purpose of this study was to explore the APS-R and FAPS for Asian Indians and examine the relationship between several of the constructs related to perfectionism and conduct a latent profile analysis	Indian	Survey	Discussion of Indian culture and collectivist values	MGCFA, hierarchical regression

Table 13: Configural Analysis Results

	Group 1	**FGCS**			**Group 2**	**CGCS**		
	Estimate	Std.Err	z-value	est.std	Estimate	Std.Err	z-value	est.std
RIL								
RIintegrate	1			0.72420091	1			0.722965
RIsocietal	1.035	0.021	48.819	0.749434482	1.009	0.022	45.254	0.729754
RIdiverse	1.025	0.023	43.855	0.742390008	0.953	0.024	40.176	0.689225
RIownview	1.086	0.023	47.467	0.78679549	1.06	0.023	45.296	0.766363
RIperspect	1.05	0.024	44.58	0.760254289	0.992	0.025	40.027	0.717439
RInewview	1.098	0.024	46.312	0.794974723	1.051	0.024	43.607	0.75974
RIconnect	1.106	0.023	48.447	0.800892449	1.085	0.024	45.044	0.784634
HOL								
HOapply	1			0.782425526	1			0.728082
HOanalyze	1.106	0.017	63.222	0.865241135	1.095	0.023	47.644	0.797364
HOevaluate	1.113	0.018	60.546	0.871154294	1.096	0.026	41.873	0.798023
HOform	1.055	0.018	57.133	0.825806274	1.117	0.025	44.478	0.813414
QuantR								
QRconclude	1			0.761647395	1			0.733609
QRproblem	1.126	0.025	45.794	0.857827903	1.161	0.028	40.825	0.85157
QRevaluate	1.151	0.025	45.783	0.87656869	1.159	0.029	39.817	0.850011
LearnS								
LSreading	1			0.82108648	1			0.823296
LSnotes	0.929	0.027	34.449	0.76249458	0.889	0.025	35.545	0.732305
LSsummary	0.996	0.027	36.684	0.817474474	0.97	0.026	37.121	0.79837
ColL								

	Group 1	**FGCS**			**Group 2**	**CGCS**		
	Estimate	Std.Err	z-value	est.std	Estimate	Std.Err	z-value	est.std
CLaskhelp	1			0.687016204	1			0.725282
CLexplain	1.168	0.033	35.352	0.802316924	1.126	0.03	38.149	0.816923
CLstudy	1.206	0.032	37.933	0.828300065	1.107	0.027	40.461	0.802995
CLproject	1.088	0.033	33.011	0.747563815	1.083	0.028	39.123	0.785711
DivD								
DDrace	1			0.871545956	1			0.839636
DDeconomic	1.043	0.012	89.221	0.909194289	1.087	0.016	66.044	0.913027
DDreligion	1.024	0.011	95.724	0.892149151	0.99	0.014	71.613	0.830881
DDpolitical	0.939	0.013	74.928	0.818464632	0.892	0.017	53.256	0.749178
SFI								
SFcareer	1			0.792659089	1			0.749839
SFotherwork	0.975	0.025	39.342	0.772617684	1.029	0.031	33.569	0.771262
SFdiscuss	1.072	0.025	43.453	0.849736446	1.023	0.03	33.96	0.766851
SFperform	1.022	0.024	42.771	0.810468067	1.006	0.03	33.838	0.754168
EffTF								
ETgoals	1			0.798556945	1			0.73349
ETorganize	0.968	0.021	46.761	0.773172553	0.983	0.025	38.763	0.72067
ETexample	1.027	0.022	46.497	0.820032962	1.08	0.026	40.908	0.792015
ETdraftfb	0.986	0.021	47.353	0.787543122	0.977	0.027	36.474	0.716372
ETfeedback	1.01	0.021	47.877	0.806936341	1.039	0.028	37.241	0.761956
QualI								
QIstudent	1			0.685011535	1			0.631419
QIadvisor	1.091	0.029	37.708	0.747062729	1.03	0.035	29.064	0.650558
QIfaculty	1.208	0.031	39.545	0.82778373	1.272	0.039	32.371	0.802912
QIstaff	1.073	0.028	37.662	0.73503391	1.078	0.037	29.375	0.680741
QIadmin	1.072	0.028	37.625	0.734333858	1.053	0.036	28.886	0.665007
SupEv								

	Group 1	**FGCS**			**Group 2**	**CGCS**		
	Estimate	Std.Err	z-value	est.std	Estimate	Std.Err	z-value	est.std
SEacademic	1			0.8008285	1			0.783521
SElearnsup	0.94	0.016	60.268	0.75269515	0.912	0.018	50.309	0.714575
SEdiverse	1.005	0.015	65.07	0.805169958	0.953	0.018	53.009	0.746507
SEsocial	1.079	0.014	74.887	0.864298354	1.05	0.017	60.968	0.822358
SEwellness	1.045	0.015	69.236	0.836571961	1.004	0.018	56.368	0.786538
SEnonacad	0.934	0.016	56.951	0.748234091	0.884	0.02	43.819	0.692645
SEactivities	1.006	0.015	67.208	0.805843244	0.962	0.019	50.742	0.753707
SEevents	1.013	0.016	64.863	0.811385131	0.994	0.019	52.371	0.77856

Table 14: Metric Analysis Results

	Group 1	**FGCS**			**Group 2**	**CGCS**		
	Estimate	Std.Err	z-value	est.std	Estimate	Std.Err	z-value	est.std
RIL								
RIintegrate	1			0.734661	1			0.708686
RIsocietal	1.024	0.015	66.381	0.752373	1.024	0.015	66.381	0.725771
RIdiverse	0.995	0.017	59.288	0.731189	0.995	0.017	59.288	0.705337
RIownview	1.075	0.016	65.176	0.790037	1.075	0.016	65.176	0.762104
RIperspect	1.026	0.017	59.878	0.75366	1.026	0.017	59.878	0.727013
RInewview	1.078	0.017	63.287	0.792131	1.078	0.017	63.287	0.764124
RIconnect	1.097	0.017	65.907	0.80589	1.097	0.017	65.907	0.777396
HOL								
HOapply	1			0.780453	1			0.731692
HOanalyze	1.102	0.014	79.124	0.860049	1.102	0.014	79.124	0.806314
HOevaluate	1.107	0.015	73.115	0.863706	1.107	0.015	73.115	0.809743

	Group 1	**FGCS**			**Group 2**	**CGCS**		
	Estimate	Std.Err	z-value	est.std	Estimate	Std.Err	z-value	est.std
HOform	1.079	0.015	72.457	0.841937	1.079	0.015	72.457	0.789334
QuantR								
QRconclude	1			0.757263	1			0.739959
QRproblem	1.14	0.019	61.069	0.86319	1.14	0.019	61.069	0.843465
QRevaluate	1.154	0.019	60.489	0.87383	1.154	0.019	60.489	0.853862
LearnS								
LSreading	1			0.829748	1			0.811787
LSnotes	0.912	0.019	48.727	0.75671	0.912	0.019	48.727	0.740331
LSsummary	0.985	0.019	51.453	0.81709	0.985	0.019	51.453	0.799404
ColL								
CLaskhelp	1			0.698068	1			0.711433
CLexplain	1.149	0.022	51.128	0.80213	1.149	0.022	51.128	0.817487
CLstudy	1.161	0.021	54.78	0.810541	1.161	0.021	54.78	0.826059
CLproject	1.085	0.022	49.568	0.757534	1.085	0.022	49.568	0.772037
DivD								
DDrace	1			0.873127	1			0.838314
DDeconomic	1.06	0.01	110.985	0.925289	1.06	0.01	110.985	0.888396
DDreligion	1.01	0.008	119.472	0.8821	1.01	0.008	119.472	0.846929
DDpolitical	0.921	0.01	91.611	0.804239	0.921	0.01	91.611	0.772172
SFI								
SFcareer	1			0.79438	1			0.747828
SFotherwork	0.995	0.019	51.558	0.790329	0.995	0.019	51.558	0.744013
SFdiscuss	1.053	0.019	55.181	0.83619	1.053	0.019	55.181	0.787187
SFperform	1.016	0.019	54.533	0.80732	1.016	0.019	54.533	0.760009
EffTF								
ETgoals	1			0.792804	1			0.741348

	Group 1	**FGCS**			**Group 2**	**CGCS**		
	Estimate	Std.Err	z-value	est.std	Estimate	Std.Err	z-value	est.std
ETorganize	0.974	0.016	60.69	0.772023	0.974	0.016	60.69	0.721916
ETexample	1.048	0.017	61.798	0.830951	1.048	0.017	61.798	0.777019
ETdraftfb	0.982	0.016	59.734	0.778922	0.982	0.016	59.734	0.728367
ETfeedback	1.022	0.017	60.574	0.810407	1.022	0.017	60.574	0.757809
QualI								
QIstudent	1			0.684943	1			0.631451
QIadvisor	1.068	0.022	47.705	0.731462	1.068	0.022	47.705	0.674337
QIfaculty	1.232	0.024	50.996	0.843941	1.232	0.024	50.996	0.778032
QIstaff	1.075	0.023	47.766	0.736294	1.075	0.023	47.766	0.678791
QIadmin	1.065	0.022	47.439	0.729387	1.065	0.022	47.439	0.672424
SupEv								
SEacademic	1			0.811555	1			0.768063
SElearnsup	0.928	0.012	78.487	0.75352	0.928	0.012	78.487	0.713138
SEdiverse	0.984	0.012	83.97	0.798754	0.984	0.012	83.97	0.755948
SEsocial	1.067	0.011	96.637	0.866213	1.067	0.011	96.637	0.819791
SEwellness	1.028	0.012	89.342	0.834462	1.028	0.012	89.342	0.789741
SEnonacad	0.914	0.013	71.809	0.741984	0.914	0.013	71.809	0.702219
SEactivities	0.988	0.012	84.134	0.802128	0.988	0.012	84.134	0.75914
SEevents	1.005	0.012	83.391	0.815889	1.005	0.012	83.391	0.772164

Table 15: Scalar Analysis Results

	Group 1	**FGCS**			**Group 2**	**CGCS**		
	Estimate	Std.Err	z-value	est.std	Estimate	Std.Err	z-value	est.std
RIL								
RIintegrate	1			0.7271	1			0.719073
RIsocietal	1.031	0.019	53.351	0.749669	1.031	0.019	53.351	0.729452
RIdiverse	1.018	0.021	47.746	0.7401	1.018	0.021	47.746	0.692639
RIownview	1.083	0.021	51.804	0.787319	1.083	0.021	51.804	0.765708
RIperspect	1.043	0.021	48.624	0.758508	1.043	0.021	48.624	0.719935
RInewview	1.095	0.022	50.556	0.795826	1.095	0.022	50.556	0.758752
RIconnect	1.101	0.021	52.55	0.800756	1.101	0.021	52.55	0.784707
HOL								
HOapply	1			0.779941	1			0.732004
HOanalyze	1.107	0.016	67.345	0.863357	1.107	0.016	67.345	0.800441
HOevaluate	1.118	0.017	64.447	0.87182	1.118	0.017	64.447	0.797005
HOform	1.063	0.017	61.143	0.829125	1.063	0.017	61.143	0.808426
QuantR								
QRconclude	1			0.763234	1			0.731923
QRproblem	1.124	0.023	49.859	0.857544	1.124	0.023	49.859	0.851776
QRevaluate	1.148	0.023	49.819	0.875867	1.148	0.023	49.819	0.850945
LearnS								
LSreading	1			0.822314	1			0.821706
LSnotes	0.927	0.025	37.784	0.762696	0.927	0.025	37.784	0.731987
LSsummary	0.993	0.025	40.273	0.816414	0.993	0.025	40.273	0.799883
ColL								
CLaskhelp	1			0.688249	1			0.722889
CLexplain	1.165	0.03	39.432	0.801895	1.165	0.03	39.432	0.81886
CLstudy	1.2	0.029	42.057	0.826024	1.2	0.029	42.057	0.806127

	Group 1	**FGCS**			**Group 2**	**CGCS**		
	Estimate	Std.Err	z-value	est.std	Estimate	Std.Err	z-value	est.std
CLproject	1.089	0.029	37.108	0.749438	1.089	0.029	37.108	0.782557
DivD								
DDrace	1			0.873394	1			0.836282
DDeconomic	1.042	0.011	93.508	0.910335	1.042	0.011	93.508	0.91108
DDreligion	1.02	0.01	100.656	0.890914	1.02	0.01	100.656	0.833255
DDpolitical	0.934	0.012	78.82	0.815923	0.934	0.012	78.82	0.75452
SFI								
SFcareer	1			0.789906	1			0.754308
SFotherwork	0.986	0.023	43.006	0.779211	0.986	0.023	43.006	0.760572
SFdiscuss	1.074	0.023	47.124	0.848563	1.074	0.023	47.124	0.767865
SFperform	1.023	0.022	46.973	0.808084	1.023	0.022	46.973	0.758662
EffTF								
ETgoals	1			0.796328	1			0.736037
ETorganize	0.967	0.019	50.758	0.770023	0.967	0.019	50.758	0.724558
ETexample	1.03	0.021	50.122	0.820054	1.03	0.021	50.122	0.791482
ETdraftfb	0.993	0.019	51.565	0.79047	0.993	0.019	51.565	0.713572
ETfeedback	1.016	0.019	52.3	0.809251	1.016	0.019	52.3	0.758939
QualI								
QIstudent	1			0.687994	1			0.628014
QIadvisor	1.084	0.024	44.596	0.745546	1.084	0.024	44.596	0.652061
QIfaculty	1.198	0.026	46.58	0.824371	1.198	0.026	46.58	0.80799
QIstaff	1.066	0.024	44.419	0.733186	1.066	0.024	44.419	0.682483
QIadmin	1.075	0.024	44.523	0.739613	1.075	0.024	44.523	0.65864
SupEv								
SEacademic	1			0.802424	1			0.781162
SElearnsup	0.94	0.015	64.802	0.75446	0.94	0.015	64.802	0.711893

	Group 1	**FGCS**			**Group 2**	**CGCS**		
	Estimate	Std.Err	z-value	est.std	Estimate	Std.Err	z-value	est.std
SEdiverse	1.008	0.014	69.848	0.808966	1.008	0.014	69.848	0.740598
SEsocial	1.078	0.013	80.067	0.86485	1.078	0.013	80.067	0.821397
SEwellness	1.04	0.014	74.122	0.834736	1.04	0.014	74.122	0.789187
SEnonacad	0.931	0.015	60.927	0.746739	0.931	0.015	60.927	0.694957
SEactivities	1	0.014	71.64	0.80219	1	0.014	71.64	0.759388
SEevents	1.011	0.015	69.412	0.811089	1.011	0.015	69.412	0.779445