MOTEANE, ‘MALITŠITSO. Ph.D. Critically exploring the use of race and ethnicity as grouping variables in studies that use or include differential item functioning analyses. (2024)
Directed by Dr. Micheline Chalhoub-Deville. 215 pp.

This study addresses the underexplored terrain of conceptualizing and operationalizing race and ethnicity as grouping variables in Differential Item Functioning (DIF) studies within the context of psychometric research. The investigation extends beyond the identification of DIF and delves into the theoretical framing and communication of findings related to these variables. Analyzing 120 articles from diverse academic databases, this research employs descriptive statistics and interviews with two authors in its two-phase mixed methods approach.

The results illuminate significant gaps in the current practices of DIF studies utilizing race and ethnicity as grouping variables. Notably, 75% of studies need more operational definitions and theoretical justifications for the inclusion of the race and ethnicity variables. The diversity in the definitions employed, often aligning with census categories, and the varied approaches to participant categorization (57% allowing self-selection, 30% unspecified) underscore the need for methodological clarity. The prevalence of an exploratory approach (83%) to DIF detection, with a limited focus (29%) on threats to internal validity, indicates a nuanced landscape.

In conclusion, this study highlights the complexity surrounding the use of race and ethnicity as grouping variables in DIF studies. It emphasizes the necessity for clearer conceptualization, theoretical framing, and interpretation of findings. It advocates for enhanced methodological rigor, transparency, and cross-cultural considerations in psychometric research, paving the way for more nuanced and reliable assessments of differential item functioning across diverse populations.

CRITICALLY EXPLORING THE USE OF RACE AND ETHNICITY AS GROUPING VARIABLES IN STUDIES THAT USE OR INCLUDE DIFFERENTIAL ITEM FUNCTIONING ANALYSES

by

‘Malitšitso Moteane

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

Greensboro

2024

Approved by

Dr. Micheline Chalhoub-Deville
Committee Chair

DEDICATION

I dedicate this dissertation to the brightest star in the constellation that is my life, Monono Pabala Thahane. For being brave and crossing the Atlantic Ocean to come on this adventure with me, for being who I have always wished I was; confident, articulate, fabulous, and for being a constant reminder that I do indeed matter; accept this offering as a thank you Mohlakoana oa ‘Mapholo oa Lisema, maila ngoatheloa. Motho a sa jeng sengoathoana sa maobane, a jang polokoe kaofela.

I dedicate this also to my village, Mme Advocate Tiisetso Mafatle; I cannot count the ways in which this achievement is as much yours as it is mine, my best friends Lineo Tsikoane and Mpoetsi Mabathoana for sending me off and holding me in your thoughts and prayers, and my brothers, sisters, cousins maternal and paternal, for literally giving of yourselves to pay for the editing of this manuscript and being unwavering in your faith in me thank you, thank you, thank you.

APPROVAL PAGE

This dissertation written by ‘Malitšitso Moteane has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair

Dr. Micheline Chalhoub-Deville

Committee Members

Dr. Devdass Sunnassee

Dr. Ye He

March 18, 2024

Date of Acceptance by Committee

February 12, 2024

Date of Final Oral Examination

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my dissertation committee chair, Dr. Chalhoub-Deville, for her unwavering support, invaluable guidance, and mentorship throughout this journey. Her encouragement to embrace courage and amplify my voice has been instrumental in shaping this work.

I am also grateful to my dissertation committee members: Dr. He for her expertise in methodological development and her continual encouragement and insightful advice, which have been pivotal in refining the approach of this research and Dr. Sunnassee for his meticulous feedback and constant encouragement, which have greatly enriched the quality of this dissertation.

I extend my heartfelt appreciation to Dr. Henson for his role as a critical friend and both formal and informal reflective practice partner, contributing significantly to the depth and rigor of this study.

Lastly, I want to thank Dr. Brianna Hooks-Singletary for holding me to the highest standard of quality and forbidding me from doing a simulation study. Thank you, my friend, for being a great listener, a thought partner, and my biggest cheerleader. I am forever indebted to you. I love you.

Their collective contributions have been indispensable in shaping this work, and I am profoundly grateful for their support and guidance throughout this academic endeavor.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I: INTRODUCTION

Differential item functioning (DIF) analyses are often conducted during test development and test scoring as a precursor to determining whether certain items on a test are biased against one or more manifest groups of test takers. There is a long tradition in measurement research, in part motivated by the protection of minoritized racial and ethnic groups after the passage of the Civil Rights Act of 1964, of using race and/or ethnicity as grouping variables, and there are generally accepted norms of how groups are referred to and compared. Racial and ethnic groups and groupings are highly fluid over time and geography and can often only be understood in the political and social contexts within which they are used. It is, therefore, important to understand how they are employed as grouping variables in DIF studies as no prior studies have explored, critically or otherwise, how the race and ethnicity variables are conceptualized, operationalized, and used in such studies.

The conjoined growth of psychometrics and eugenics in the late 19th and early 20th century, wherein the former was used as the scientific basis for the latter, referred to by psychometrists as the "dark ages" of the field (Rust et al., 2020; Wijsen & Borsboom, 2021), is a big part of the interrelationship between grouping variables, especially race and ethnicity and psychometric techniques such as DIF.

To critically examine how race and ethnicity are used as grouping variables in studies of DIF, a sequential explanatory mixed methods design was used. In the first quantitative stage of the study, published DIF studies that use race and/or ethnicity as grouping variables were collected and systematically reviewed to uncover trends in how the variables are used, the type of DIF detection methods employed and the nature of reporting and interpreting findings of such studies. The findings from this phase were then used to identify a purposive sample to further

explore, through qualitative interviews, the rationale, and mechanics of using the race and ethnicity grouping variables and a reflection on the interpretation and possible uses of research findings of DIF studies.

This study looked across articles that included DIF analyses where race and/or ethnicity were used as grouping variables to uncover trends in how race and ethnicity are conceptualized, operationalized, and used. Current literature on DIF studies focuses on the technical features of different DIF detection methods, including their accuracy at detecting uniform and non-uniform DIF using both real and simulated data sets. This study provides insight into how the race and/or ethnicity variable is used, which will inform how DIF related to those grouping variables is simulated. In addition, the in-depth insights from researchers that published DIF studies will contribute to what is understood about the race and ethnicity variables as they are used in DIF studies, which can, when discussed from the Critical Race Theory (CRT) lens, provide a collection of factors for researchers to consider when using these variables in DIF studies.

Background and Context

Psychological tests administered to groups of test-takers in the United States are recorded to have taken off in earnest in the form of Army Alpha and Army Beta tests in the early 20th century in support of World War 1 (Aiken, 1985; Anastasi, 1976; Valencia, 1997) following the adaptation of the Stanford-Binet intelligence tests and objective test items that lent themselves to administration to large groups of test takers (Anastasi, 1976). In addition to being used for administrative decisions within the army, such as admission into service, duty assignment and discharge from service, intelligence (later aptitude) tests became widely used by the public for various purposes that still subsist to this day (Anastasi, 1976). For instance, education decisions

for employment purposes and college admissions continue to be based, at least in part, on test scores.

Academic attention to bias and fairness in testing proliferated in the 1960's and 1970's in response to the changing political climate from one where African Americans and other minoritized groups were discriminated against on the basis of their race to one where such discrimination was illegal (Angoff, 1993; Cole, 1993; Jonson & Geisinger, 2022). More and more technical approaches to the detection of bias were put forth, which ultimately led to the distinction between the statistically observable component, DIF, and the substantive component which is an articulation of the reason for the differential functioning and a determination whether that reason is relevant or irrelevant to the construct being measured. This approach was intended to disentangle DIF from the loaded political and social connotations of the term bias (Angoff, 1993; Zumbo, 1999).

In the application of DIF analyses, it is customary to distinguish between the groups being considered by referring to the group that is hypothesized to the be advantaged by the item as the reference group and the group that is hypothesized to be disadvantaged by an item as the focal group (de Ayala, 2009; Osterlind & Everson, 2009; Shultz et al., 2014). The tradition has been to assign majority populations (e.g., males, and Caucasians) as the reference group while minoritized populations (e.g., females, Blacks and Hispanics) to the focal group (e.g. de Ayala, 2009; Osterlind & Everson, 2009).

Race and ethnicity are widely accepted as social constructions (Kivisto & Croll, 2012) not rooted in any biological differences that, notwithstanding, greatly influence if not predetermine how benefits accrue to individuals in space and time. The mutable nature of race

and ethnicity warrant investigation when used as grouping variable in the application of analyses such as DIF.

Theoretical Framework

This study was framed by two adjacent derivatives of Critical Race Theory (CRT): critical race quantitative intersectionality (CRQI) (Covarrubias & Vélez, 2013) and QuantCrit (Gillborn et al., 2018). Both put forth considerations to be made when applying the tenets of CRT to quantitative studies. There is a significant overlap between the two frameworks. One of the foundational concepts is the notion that race and racism, though not always explicitly addressed in quantitative research, are ever-present. In addition, critical quantitative researchers believe that statistical analysis can and should play a role in the struggles for social justice. Critical quantitative researchers insist that data cannot speak for itself and that marginalized populations' experiences, knowledge, and insight of should inform critical analyses.

Numbers are not objective or neutral and have, in the past, been used to serve deficit characterizations of Black people and other minoritized populations to serve white political and racial interests. Critical quantitative researchers also acknowledge that the categories used in traditional quantitative research are neither natural nor given and cannot, as is sometimes the case, be the cause of patterns or differences.

Critical race quantitative intersectionality explicitly states that the typical binaries (male, female) and essentializing categories (racial/ethnic groupings) ignore the unique experiences of those at the intersections. Thus, they suggest expanding grouping variables to enrich findings from quantitative analyses.

Problem Statement

Validity is conceived as a unitary concept, and multiple sources of evidence are required to validate the intended uses of the test (American Educational Research Association et al., 2014; Kane, 2006; Messick, 1989). Five sources of valid evidence are outlined in the Standards for Educational and Psychological Testing (American Educational Research Association et al., 2014): evidence based on test content, evidence based on response processes, evidence based on internal structure, evidence based on relations to other variables and evidence based on the consequences of testing (American Educational Research Association et al., 2014). Validation is the process of investigating the validity of the use of test score interpretations for their intended use and as conceptualized through an Interpretation and Use Argument (IUA) (Kane, 2013) that organizes evidence from the five sources to evaluate their appropriateness or defensibility.

Fairness in testing is an integral part of validity that should be attended to at all stages of test development (American Educational Research Association et al., 2014). Fairness can manifest in several ways, such as the treatment of examinees during a test administration, test instructions or test formats (such as computer-administered tests) that affect some examinees more than others or measurement bias.

Measurement bias can be thought of as any systematic error in test scores. Items systematically favoring one demographic group over another (displaying differential item functioning, DIF) are one form of measurement bias. Showing that test items function similarly for different groups of examinees (measurement invariance) is another form of evidence based on the internal structure evidence (Lane, 1999).

Differential Item Functioning (DIF)

Test items display statistical DIF when test takers with the same standing on a construct but belonging to different manifest groups **systematically** have different probabilities of endorsing the correct answer (Angoff, 1993; Clauser & Mazor, 1998; Osterlind & Everson, 2009; Roussos & Stout, 2004). DIF can also be understood as an unexpected multidimensionality of the test item in question. If, upon further investigation, it is found that the cause of DIF is unrelated to the primary construct being measured, then the item is said to display measurement bias (Zumbo, 1999).

At its core, DIF is concerned with differing response patterns of test takers as a function of their group membership. While there are several ways to form the groups to be compared in DIF analysis, such as considering course-taking behavior (Bandalos, 2018), instructional background or test-wiseness (Roussos & Stout, 2004), groups are normally formed along demographic lines. Often, DIF analysis are centered around comparisons between Whites and ethnic and racial groups protected by Civil Rights laws (Angoff, 1993). Race and ethnicity are often used as grouping variables in DIF studies, but how these variables are conceptualized and operationalized has yet to be studied in the context of psychometric research. Further, how the findings of such studies are interpreted and communicated by the researcher and by the readership is also unexamined.

Purpose of the Study

This study aimed to explore trends in the conceptualization of race and ethnicity as grouping variables, the theoretical framing, interpretation, and the onward use of findings in differential item functioning (DIF) studies that use race/ethnicity as a grouping variable. An explanatory sequential mixed methods design involved collecting quantitative data and then

explaining the quantitative results with in-depth qualitative data. In the first quantitative strand of the study, data was collected from DIF studies published in peer-reviewed academic journals that include race and/or ethnicity as grouping variables. The second qualitative phase was conducted as a follow-up to the quantitative results and to help explain researchers' conception and rationalization of race and ethnicity as a grouping variable and to explore whether their findings, as used in forward citations, are interpreted in alignment with their intentions at the time of publication. In this explanatory follow-up phase, the critical race theory (CRT) was used to explain how a purposeful sample of researchers with published DIF studies in peer-reviewed journals justify and operationalize race and/or ethnicity as grouping variables and interpret their findings and onward citations.

Research Questions

The following research questions will guide this mixed methods study:

RQ1: What are key characteristics of the differential item functioning investigations that employ race and/or ethnicity as grouping variables and appear in peer-reviewed journal articles published in 2015-2020 literature? [Journals, number of authors, author affiliations, fields of study, DIF detection methods, theoretical framing of DIF, definitions of DIF/operationalization]

RQ2: What trends emerge in differential item functioning (DIF) analyses reported in recently published research studies that use race and/or ethnicity as grouping variables in terms of

a) how the terms are defined,

b) how the categorization is conducted, and

c) how findings are reported and interpreted?

RQ3: How does the conceptualization of race and ethnicity in recently published DIF studies by the researchers identified for RQ1 respond to the particularities of the study contexts?

RQ4: To what extent do findings from DIF studies that use race and/or ethnicity as grouping variables align with the authors they cite in terms of interpretation when utilized in future research?

Significance of the Study

This study will add to systematic reviews of DIF simulation studies to provide a broader picture of the reporting of DIF studies in peer-reviewed academic journals. Systematic literature review and syntheses have focused on refining DIF detection methods, simulation studies, bibliometric features, and field-specific (language testing) synthesis.

Methodologically, the study will also provide an overview of how the race and ethnicity variables are used in studies that include DIF and what insights the use of the variable provides for researchers. Including a qualitative phase, which gave researchers a voice to reflect on whether and to what extent their reported findings of DIF studies are used in the ways that they intended, is novel in psychometric research. In addition, the use of forward citations as an interview prompt is new and offers an avenue for researchers to see and reflect on the impact of their work in their field. It is further hoped that the findings from this study will also provide insight into things to consider for future researchers when conducting DIF studies.

Nature of the Study

An explanatory sequential mixed methods study was employed to address the research questions listed above. The first phase of the study was a systematic literature review of articles of DIF that use race and/or ethnicity as grouping variable to uncover trends in the conceptualization of race and ethnicity, the theoretical framing of the studies, the findings of the

studies and how the results are interpreted in relation to the grouping variables. The findings from this quantitative phase will be used to select cases for the second qualitative phase of the study. Extreme case sampling was used to identify US-based articles where DIF research findings were fully interpreted, including mention of the grouping variable and the direction of DIF presented or fully interpreted and situated in the broader sociopolitical context. Authors of those studies were invited for an interview to reflect on their study and some forward citations of their work.

Limitations

The following limitations/delimitations are present in this study:

- Location bias. It is possible that some research journals which also publish DIF studies may not be indexed by the five bibliographic databases included in the study. While the choice of databases is informed by other systematic reviews related to DIF and the availability of said databases through the University of North Carolina at Greensboro (UNCG) Library, some studies may be excluded for this reason. In addition, many DIF studies are conducted by testing companies (e.g., Educational Testing Services (ETS)) , recorded as either internal or external reports, and not necessarily published in peer-reviewed journals; thus, the study excludes such studies from consideration.
- Multiple publication bias. Some manuscripts may refer to the same study, which may result in an overrepresentation of study qualities and framing, especially in the quantitative phase of the study, where the research publication is the unit of analysis.

- The qualitative phase focuses only on DIF researchers. The design of the study invites DIF researchers to reflect on how the findings of their studies are used in forward citations, particularly concerning the race and/or ethnicity variables, by presenting them with forward citations. This design does not incorporate the interpretation of the researchers that cite articles beyond the citation, which would provide a deeper understanding of how DIF findings are perceived and used.
- My identity as a Black female graduate student could potentially delimit the data that can be collected from researchers who use DIF analyses in their published research papers to the extent that their race/ethnicity and gender are different to my own, primarily white males. While I view my positionality as a strength, a well-established feature of qualitative inquiry and to be in great alignment with Critical Race Theory as it draws on the unique knowledge and insights of Black and Brown scholars as an invaluable asset, I also acknowledge and recognize that tensions arise when engaging in mixed methods between the objective stance of but a challenge

CHAPTER II: LITERATURE REVIEW

This chapter presents the literature relevant to my research problem and is organized as follows. The chapter begins with a presentation of the theoretical framework that will guide the research, after which a brief history of psychometrics as a field will serve as the backdrop to introducing studies of item bias, which later became studies of differential item functioning (DIF). The chapter will then introduce race and ethnicity, particularly in the context of research, and conclude with how race and ethnicity are presented in psychometrics as articulated in the Standards of Educational and Psychological Testing.

Theoretical Framework

This study will be framed by two adjacent derivatives of Critical Race Theory (CRT): critical race quantitative intersectionality – CRQI (Covarrubias et al, 2013; 2018) and QuantCrit (Gillborn et al., 2018). I begin this section by describing CRT and some of its central tenets and tracing the history of the use of CRT in educational research. I will then link the central tenets of CRT to CRQI and QuantCrit as applicable and conclude by highlighting the tenets that inform the research questions and the methodology, specifically the data analysis.

Critical Race Theory, Critical Race Quantitative Intersectionality, and QuantCrit

Critical Race Theory (CRT) is traced back to the work of black critical legal scholars from the 1970s onwards (Delgado & Stefancic, 2017). The founders of CRT, at a workshop held in 1989 in Wisconsin of critical legal scholars of color, were dissatisfied with the failure of critical legal studies to address the lived experiences of people of color (especially Blacks) in the face of the law articulated the founding tenets of the theory (Delgado & Stefancic, 2021). They held that race and racism were central to understanding the mechanisms that permeate all of life

and lived experiences in the United States of America (Delgado & Stefancic, 2017; Dixson & Rousseau, 2005). The call for and subsequent application of CRT in education is credited to seminal papers by William Tate and Gloria Ladson-Billings in the mid-1990s (Ladson-Billings & Tate, 1995; Tate, 1994). This call to center race and racism in the consideration of the past and current state of education in the USA stemmed from their argued link between race (particularly whiteness) and property, which logically results in starkly different schooling experiences for students who are raced differently (Ladson-Billings & Tate, 1995).

Critical Race Quantitative Intersectionality (CQRI) and QuantCrit are frameworks or considerations to be made when applying the tenets of CRT to quantitative studies. CRQI was articulated as a chapter in the Handbook of Critical Race Theory in Education (Covarrubias & Vélez, 2013). QuantCrit debuted in a special Race, Ethnicity, and Education issue in 2018.

In the following section, I outline CRT tenets articulated by scholars that have been applied specifically to quantitative research methods in CRQI and QuantCrit. A summary of this mapping of CRT tenets to CRQI and QuantCrit is provided in Table 1.

The first is that ***racism is ordinary, not aberrant*** and is woven into the daily lives of Americans through systems and institutions (Delgado & Stefancic, 2017). This endemic nature of racism, critical race scholars theorize, makes it challenging to assail as it hides in plain sight. QuantCrit scholars articulate this tenet as the necessary and critical departure point for their work. They center the non-neutrality of researchers and research and invoke the 'critical race-conscious perspective' (Gillborn et al., 2018, p. 169) to ensure that quantitative research does not perpetuate racial and social inequities.

The second tenet of CRT, often referred to as ***interest convergence***, posits that any gains in the struggle for equity for minoritized and marginalized groups are only realized when they

converge with the interests of Whites. The classic example offered by legal scholars such as Derrick Bell of this tenet is a reinterpretation of the 1954 Brown v. Board of Education judgement as a revamping of the appearance and credibility of America as a non-Communist state, transitioning the southern states into industrialization which could not co-exist with segregation (Brown & Jackson, 2014, p. 16; Delgado & Stefancic, 2021, p. 24). This tenet is not mapped explicitly to CRQI or QuantCrit. However this retrospective look at supposed gains in the struggle for social justice is the impetus for maintaining a critical eye on policy and research findings to see whose interests they serve even as they proclaim to be in tandem with the pursuit of social equity.

Intersectionality is the third tenet of CRT discussed in this section. It relates to the unique experiences of those who occupy the intersection of known sites of oppression (Delgado & Stefancic, 2017, p. 58), such as race, gender, sexual orientation, ability, language proficiency, and so on. This tenet, translated to CRQI, holds that broad categorizations along race, gender, sexual orientation, and ability lines, in addition to posing a measurement challenge, homogenize largely heterogeneous groups. This hides identifiable variability within groups at the intersection of such broad categories (Covarrubias & Vélez, 2013, p. 275). Indeed, the impetus for the CRQI framework is to explore and quantify the material impact of the intersectionality of race and racism and other forms of oppression and subordination (Covarrubias & Vélez, 2013). QuantCrit does not articulate intersectionality or intersectional data mining as an explicit tenet.

CRT also posits that contemporary ***research is not neutral*** and that researchers bring their lived experiences, biases, and assumptions to the practice of research. This opposition to mainstream notions of research is intended to apply to all its aspects, including the questions , the data sources identified, the tools for collecting said data, how data are analyzed and how results

are interpreted and disseminated. CRT scholars argue that current ideas about race and people of color were informed by scientific research and that the taken-for-granted assumptions, data collection, and analysis tools and procedures should be made explicit and interrogated (Covarrubias et al., 2018, p. 145). For example, the very birth of social statistics by Sir Francis Galton was used to support notions of inherent and unchanging human differences and justify discriminatory practices and racial hierarchies in the form of eugenics (Zuberi, 2001).

CRQI underscores this tenet by asserting that numbers mean nothing without their framing and that contemporary quantitative research is framed in the interests of those in power (Covarrubias & Vélez, 2013, p. 278). The objectivity of quantitative research is also questioned as data do not speak for themselves. CRQI scholarship emphasizes that it is only through the researchers' (subjective) judgement that data are collected, and a plethora of decisions made before, during and after data analysis render the process subjective. Concerning quantitative research, QuantCrit scholars hold that quantitative researchers should look behind the numbers to how they are generated, what questions are asked and how analyses are conducted to serve and normalize racial hierarchies that put Whites above Blacks and other people of color (Gillborn et al., 2018, p. 171).

CRT also holds as a central tenet that ***race is socially constructed*** as it has no genetic or biological basis (Delgado & Stefancic, 2017, p. 9). The concept of race and who falls into which category and why changes over space and time to suit the context and majoritarian interests. This CRT tenet is not mapped onto a specific CRQI tenet, as seen in Table 1. The notion that race is an unstable and highly contextual concept is expressed in seminal texts on CRQI, which underscore that the modern fields of statistics, psychology, demography, and genetics were a result of white supremacist eugenics (Covarrubias & Vélez, 2013, p. 272) and thus if used

indiscriminately run the risk of further solidifying racial hierarchies. QuantCrit takes on this tenet by reaffirming that categories are neither 'natural' nor given. It calls for a critical examination of how categories are used, operationalized, and interpreted. For instance, QuantCrit advises that when race is used to surface unequal outcomes in interpretations, we read race as racism as it is the effect of racism as opposed to a deficiency in the minoritized group (Gillborn et al., 2018, p. 171).

CRT holds that the ***voices, experiences, and insights of people of color*** are important throughout the research process (Delgado & Stefancic, 2017, pp. 11, 44–45; Dixson & Rousseau, 2005, p. 10). In the legal framing of CRT, this tenet is called counter-storytelling. It takes the form of recounting personal counter stories from the perspective of scholars of African American, Asian, Latina/o/x, or American Indian ancestry to provide insight to which white scholars or readers are not privy (Delgado & Stefancic, 2017). This tenet is reflected in both CRQI and QuantCrit. In CRQI, the role of experiential knowledge in quantitative studies is what grounds the work of CRQI scholars, from the development of research questions, methods of inquiry and analysis to the interpretation and dissemination of findings for the benefit of racialized and marginalized communities (Covarrubias & Vélez, 2013). To further illustrate, the testimonies of researchers were presented alongside quantitative findings in a CRQI + T study to buttress said findings with researchers' lived experiences s(Covarrubias et al., 2018). The author's testimonials grounded their quantitative findings and guided their analyses and interpretation. QuantCrit departs from the premise that numbers are not neutral and tend to speak for majoritarian interests, as discussed previously, and moves to center the voices and experiences of people of color to inform research, analysis, and critique (Gillborn et al., 2018, p. 173).

In its orientation, CRT is ***committed to social justice***. CRQI, as a methodological framework for quantitative research in education, orients itself as a tool for educational and social transformation (Covarrubias & Vélez, 2013, p. 280). In this vein, the intention is for the work to be collaborative with people and communities of color in response to questions asked from their perspectives that are useful to them (Covarrubias & Vélez, 2013). Similarly, QuantCrit attests that statistical analysis has no inherent value but can be coopted in the struggle for social justice.

Reconciling critical dispositions and quantitative methods, particularly CRT, which others view as antithetical, is pertinent if the social justice aims and ends of CRT scholarship are to be met (Sablan, 2014). Quantitative research occupies a privileged position both in research and policy; thus, attempts to achieve social equity need to employ this essential component of research (e.g. Garcia et al., 2018; Gillborn et al., 2018). Secondly, central to CRT is the notion that the purported neutrality of current scholarship is questionable at best and inherently racist (Dixson & Rousseau, 2005); thus, the infusion of CRT into quantitative research ensures that researchers, their values, underlying logic, and biases are explicit and not assumed to be neutral.

This section has presented the theoretical framework for this study in the form of two outgrowths of CRT, namely CRQI and QuantCrit. The two frameworks significantly overlap but differ in the number of original CRT tenets addressed and how the articulated tenets are framed. Both are relevant to the study of DIF research as they provide a framework for interrogating how the race/ethnicity variable is construed and operationalized. The following section begins with a brief history of psychometrics to position differential item functioning (DIF) before presenting an overview of DIF in testing and measurement and general research.

Table 1. Mapping of CRT Tenets to CRQI and QuantCrit

CRT		CRQI		QuantCrit
The permanence of white supremacy and racism in normal life		⇔		The centrality of racism as a deeply rooted aspect of society that is not readily amenable to quantification
Interest convergence		-		-
Intersectionality		Quantifying the Material Impact of Racism at Its Intersections: Intersectional Data Mining		-
The challenge to dominant white supremacist ideology of neutrality, objectivity, colorblindness and meritocracy	⇔	Challenging the neutrality of quantitative numbers: Data do not speak for themselves	⇔	Numbers are not neutral; they should be interrogated for their role in promoting deficit analyses that serve white racial interests

Social construction of race		⇔-		Categories/groups are neither ‘natural’ nor given and so the units and forms of analysis must be critically evaluated
Voice of color: the centrality of experiential knowledge of Black, American Indian, Asian, and Latino writers and thinkers	⇔	Originating from the experiential and material experiences of people of color	⇔	Data cannot ‘speak for itself’ and critical analyses should be informed by the experiential knowledge of marginalized groups
The commitment to a socially and racially just praxis	⇔	Being intentionally committed to addressing injustice and seeking transformation	⇔	Statistical analyses have no inherent value, but they can play a role in struggles for social justice

Historical Context of DIF

The American Psychological Association defines psychometrics as "the branch of psychology concerned with the quantification and measurement of mental attributes, behavior, performance, and the like, as well as with the design, analysis, and improvement of the tests, questionnaires, and other instruments used in such measurement" (American Psychological Association, 2023). Eugenics is defined by the National Human Genome Research Institute as "the scientifically erroneous and immoral theory of "racial improvement" and "planned breeding," which gained popularity during the early 20th century (*Eugenics and Scientific Racism*, n.d.). The links between psychometrics and eugenics have been drawn by several scholars (e.g. Dixon-Román, 2020; Jackson & Weidman, 2004; Rust et al., 2020; Valencia, 1997; Wijsen & Borsboom, 2021) through familiar influential figures in both fields such as Sir Francis Galton, Lewis Terman, Raymond Catell among others, and the fact that early intelligence tests became the scientific basis for the eugenics agenda (Rust et al., 2020; Smedley & Smedley, 2005).

While the use of psychological tests as the basis of decisions to deploy resources is said to have Chinese origins (Rust et al., 2020), many credit Sir Francis Galton and his work on the measurement of the human body and physical functions known as anthropometrics (Jones & Thissen, 2007, p. 4), and the heritability of genius (Aiken, 1985; Rust et al., 2020) in the late 1800s's. Sir Francis Galton is also credited with the coining of the term eugenics in 1883 to refer to the idea that the human race can be improved (Rust et al., 2020, p. 10; Valencia, 1997) and the ensuing policies; negative eugenics such as sterilizations or stringent immigration laws in the US (Jackson & Weidman, 2004; Valencia, 1997) Moreover, the Rassenhygiene in Nazi Germany (Jackson & Weidman, 2004; Rust et al., 2020) ensured that undesirables such as poor whites or

those of non-Nordic descent were discouraged from reproducing while “superior” people were encouraged to procreate (Jackson & Weidman, 2004).

Influences brought by James McKeen Cattell, an American psychophysicist and student of Willem Wundt in Germany, resulted in the coining of the first mental tests (Anastasi, 1976; Jones & Thissen, 2007) and birthed the idea that internal and unobservable mental processes can be measured (Dixon-Román, 2020). The development of intelligence scales in 1905 by French psychologists Alfred Binet and Theodore Simon to identify students who would benefit from remediation (Aiken, 1985; Rust et al., 2020) was taken up separately in the United States by known eugenicists H. H Goddard (Valencia, 1997) and then Stanford professor Lewis Terman. The latter resulted in the Stanford-Binet test for intelligence (Anastasi, 1976; Valencia, 1997), which was quickly repurposed when the US joined World War 1 to support the administrative decisions with measures of recruit intelligence (Anastasi, 1976; Rust et al., 2020) under the supervision of Harvard professor and then APA president Robert Yerkes (Aiken, 1985; Rust et al., 2020; Valencia, 1997). By 1918, the Army Alpha for literate and Army Beta for illiterate recruits were administered to roughly 1.7 million recruits (Rust et al., 2020; Valencia, 1997). A sample of around 10% of this large-scale administration of the Army intelligence tests, when analyzed by race with white recruits further disaggregated by nation of origin, famously found that among white test takers, those of Nordic origin scored highest and that African Americans scored the lowest which was taken to imply that they were less intelligent (Valencia, 1997).

Success with the Army tests sparked a mushrooming of psychological tests from the 1920’s. For example, the College Entrance Examinations Board (CEEB) and the American College Education (ACE) administered tests for entrance decisions into colleges; employment tests were used for entry into certain government and private sector jobs. In 1947, the testing

functions of CEEB, ACE and the Carnegie Foundation were all consolidated into Educational Testing Services (ETS) 1947 (Anastasi, 1976).

After the Second World War, geneticists and anthropologists pulled the scientific 'rug' from under the notion/concepts of racial typologies and hierarchies and the United Nations Educational, Scientific and Cultural Organization (UNESCO) put out a series of statements and/or declarations on the scientific basis for race (or rather lack thereof), race differences and racial prejudice in 1950 and 1951 (UNESCO, 1952). At around the same time, the American Psychological Association (APA) published a technical manual for psychological tests and the American Educational Research Association (AERA) and National Council of Measurement in Education (NCME) jointly produced a similar document for achievement tests in 1954 and 1955 respectively (Plake & Wise, 2014). From 1966 and roughly every 12 years after (1974, 1985, 1999, and 2014), all three bodies co-authored technical recommendations for test developers for test design, which are highly regarded by psychometricians and reflect the field's evolution. In a later section of this chapter, the latest 2014 Standards of Educational and Psychological Tests are used to trace the use of race and ethnicity as one way to decipher how the terms are conceptualized in measurement research.

Item Bias

The abridged history of the field of psychometrics provided in the previous section is described to provide the context of race and ethnicity within measurement research. This section showed the early use of tests and testing as the basis for justifying a eugenics agenda and pathologizing those of races and ethnicities other than whites. The passing of the Civil Rights Act of 1964 prohibited discrimination in public places, provided for integrating schools and other public facilities, and rendered employment discrimination illegal. The passage of the Act sparked

a flurry of research on test and item bias and techniques to detect them (Osterlind & Everson, 2009; Cole, 1993). Because standardized tests were used heavily to inform college entrance, employment, and promotion decisions, this new legislation sparked legal challenges to the tests used for these decisions. The proceedings, rulings and settlements of these legal cases reverberated in the psychometric fraternity and sparked and necessitated a more precise articulation of what constitutes bias from a psychometric perspective (Bandalos, 2018; Brown & Jackson, 2014). Bias in the social context was an ethical and moral judgment that resulted in a disproportionate allocation of resources and opportunities, but from the psychometric perspective, it was technical imprecision (Cole, 1993). It was also important to clarify terms that were the responsibility of test developers and those that were the purview of society/politics (for example, Zwick et al., (2012), Clauser & Mazor (1998)) that it is the outcomes of biased tests that created differential treatment.

DIF

This disentanglement of DIF from item bias means that the review of items for bias now occurs in two steps. The statistical part involves applying DIF detection techniques to response data to flag items. The substantive part involves flagged items then being sent for expert review on the content and construct being measured (Roussos & Stout, 2004).

Follow-up bias analyses employ content analysis, empirical evaluation, construct-related reviews and so on (Roussos & Stout, 2004; Zumbo) to determine why an item displays DIF and whether the secondary dimension is relevant to the construct being measured in a process also called “logical evidence bias” (de Ayala, 2009). If the secondary dimension is relevant to the measured construct, it is termed auxiliary DIF, interpreted as benign bias (Roussos & Stout, 2004). According to Clauser and Mazor (1998), the item is not considered biased. If the

secondary dimension is irrelevant to the measured construct(i.e., a nuisance dimension), the DIF is considered adverse (Roussos & Stout, 2004). This shift from item bias to the more palatable DIF marks what Zumbo (2007) describes as transitioning from the first generation of DIF to the second.

Technically, differential item functioning (DIF) occurs when examinees with the same standing on a construct but belonging to different manifest groups systematically have different probabilities of endorsing the correct answer (Angoff, 1993; Clauser & Mazor, 1998; Osterlind & Everson, 2009; Roussos & Stout, 2004). Table 2 provides some of the seminal definitions of DIF.

Table 2. Seminal Definitions of DIF

Authors	Definition
Holland & Thayer (1988)	"The study of items that function differently for two groups has a long history. Originally called "item bias" research, modern approaches focus on the fact that different groups of examinees may react differently to the same test question. These differences are worth exploring since they may shed light both on the test question and on the experiences and backgrounds of the different groups of examinees."
Angoff (1993)	"Differential item functioning (DIF) referring to the simple observation that an item displays different statistical properties in different group settings (after controlling for differences in the abilities of the groups)."
Clauser & Mazor (1998)	"Differential item functioning is present when examinees from different groups have differing probabilities or likelihoods of success on an item, after they have been matched on the ability of interest."

Zumbo (1999)	"DIF occurs when examinees from different groups show differing probabilities of success on (or endorsing) the item after matching on the underlying ability that the item is intended to measure."
Roussos & Stout (2004)	"DIF is said to occur in a test item when test takers of equal proficiency on the construct intended to be measured by a test, but from separate subgroups of the population, differ in their expected score on the item."
Zumbo (2007)	"DIF was the statistical term that was used to simply describe the situation in which persons from one group answered an item correctly more often than equally knowledgeable persons from another group."
Osterlind & Everson (2009)	"DIF refers to differences in the way a test item functions across demographic groups that are matched on the attribute measured by the test item."
Standards (1999)	"Differential item functioning or DIF, for short, is said to occur when examinees from groups R and F have the same degree of proficiency in a certain domain, but difference rates of success on an item. The DIF may be related to group differences in knowledge of or experience with some other topic beside the one of interest."
Standards (2014)	"For a particular item in a test, a statistical indicator of the extent to which different groups of test takers who are at the same ability level have different frequencies of correct responses or, in some cases, different rates of choosing various item options."

Statistically, for an item that is scored dichotomously, DIF occurs when the probability of an examinee from Group 1 with a proficiency of θ getting a score of 1 on item i is not equal to

the probability of a student from Group 2 with the same proficiency of θ scoring 1 on item *i*. Mathematically, for two groups, Group1 and Group 2, where Y_i is the score on a dichotomous item, *i* and θ is ability or proficiency on the construct of interest, DIF is expressed as:

$$P(\theta, Group\ 1) \neq P(Y_i = 1 \mid \theta, Group\ 2)$$

Uniform DIF occurs when one group displays a higher probability of endorsing item *i* than the other group over the entire ability/proficiency scale. Mathematically, this can be expressed as

$$P(\theta, Group\ 1) > P(\theta, Group\ 2) \quad \forall\ \theta.$$

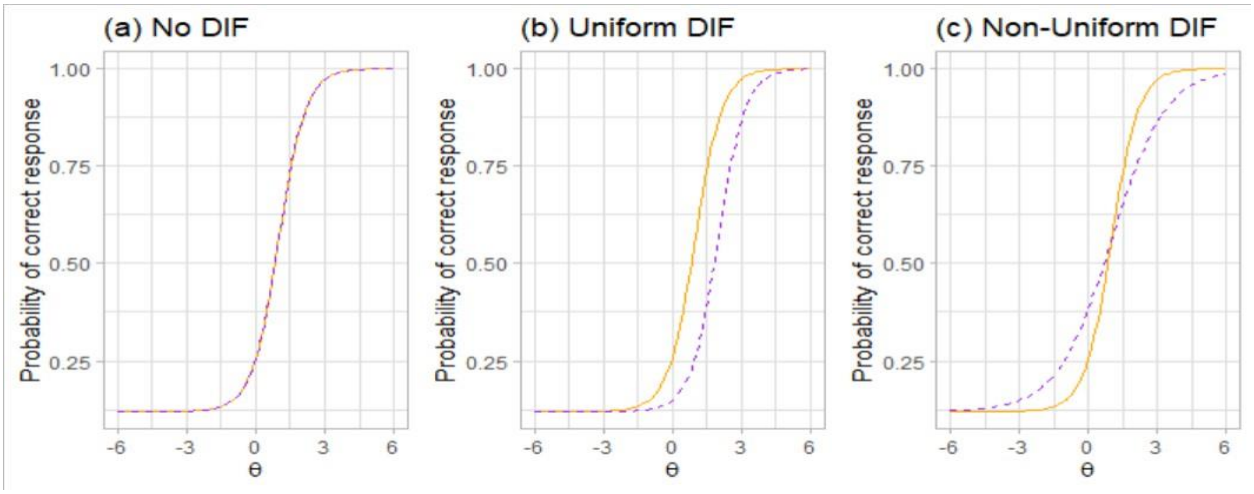
Non-uniform DIF occurs when Group 1 displays a higher probability of endorsing item *i* than Group 2 over part of the scale but a lower probability over the rest. Mathematically, for some k on the scale:

$$P(\theta, Group\ 1) > P(\theta, Group\ 2) \quad \theta < k$$

$$P(\theta, Group\ 1) < P(\theta, Group\ 2) \quad \theta \geq k$$

An Item Characteristic Curve (ICC) is a representation of the probability of a correct response to an item (on the vertical axis) at differing levels of ability/proficiency (on the horizontal axis). ICCs provide a depiction of response patterns on an item. When ICCs are drawn for the two groups on the same axis, no DIF can be visualized as the two groups' ICCs are coincident. Uniform DIF can be visualized as one ICC as a horizontal shift in the other. Non-uniform DIF is the crossing of the two groups of ICCs. Figure 1 below shows a) an item with no DIF, b) an item with uniform DIF and c) an item with non-uniform DIF.

Figure 1. ICCs Depicting no DIF, Uniform DIF and Non-Uniform DIF



Matching

A central feature of DIF detection is that examinees are matched on the proficiency, knowledge, or achievement level of the measured construct. This enables the researcher to rule out actual differences in ability as the reason for observed differences in performance by different manifest groups (Angoff, 1993). Typically, examinees are matched on their total (number correct) score. This is known as thin matching. When there are not enough examinees in each score category, and some cells of Table 2 are empty, two or more number correct scores are collapsed into score categories. This is known as thick matching.

Often, the total sum correct scores are used for matching examinees on ability; however, since it is assumed that a test measures one dimension, there are instances where section scores are more appropriate for matching (Angoff, 1993). For example, suppose a test comprises verbal and math ability when considering a verbal ability item for DIF. In that case, it is more appropriate to match examinees on the sum of verbal items correctly instead of the entire test.

It is assumed that whatever matching criterion used is valid and reliable. When this assumption is violated, and the matching criteria are invalid, superfluous DIF is detected in items

with high discrimination (Clauser & Mazor, 1998). Additionally, if the matching criterion (test) is biased, then its use in detecting DIF in the items that comprise it is problematic, to say the least (Angoff, 1993).

DIF Detection Methods

Three general classes of DIF detection methods are recorded in the literature in what Zumbo (2007) terms the second generation of DIF analyses: those based on observable statistics (non-IRT), those relying on latent IRT models, and multidimensional models. IRT and non-IRT models for DIF assume that response probabilities are a function of only one latent variable/ability (unidimensionality) (de Ayala, 2009). In practice, tests and test items are not strictly unidimensional, though unidimensional models are still useful in some cases (de Ayala, 2009), and multidimensional models better cater to this reality. The following section introduces the three classes and provides examples for each.

Non-IRT Methods

The first studies of bias, later known as differential item functioning, were conceptualized under classical test theory (CTT). They are sometimes known as observed variable methods. This section, describes two of the most popular non-IRT DIF detection methods: the Mantel Haenszel (M-H) and logistic regression.

Mantel-Haenszel

The Mantel-Haenszel method, first designed in 1959 (Angoff, 1993) for detecting the DIF is based on odd ratios that are determined by creating contingency tables. The T_k examinees who obtain the same total score, *k,* are assumed to have the same ability level and a contingency table of the same form as Table 3 below is filled. A_k is the number of students from Group 1 who get the item correct, Bk is the number of examinees from Group 1 who get the item incorrect, Ck

is the number of examinees from Group 2 who get the item correct, and finally, Dk is the number of examinees in Group who get the item incorrect.

Table 3. Contingency Table for a Studied Item for Examinees with a Total Score k

Group	1	0	Total
Group 1	A_k	B_k	N_{1k}
Group 2	C_k	D_k	N_{2k}
Total	M_{1k}	M_{2k}	T_k

The odds of getting the item correct are the probability of getting the item correct divided by the odds of missing it. In Group 1, the odds are given by $\frac{Ak}{N_{1k}} / \frac{B_k}{N_{1k}}$, while in Group 2, the odds are $\frac{Ck}{N_{2k}} / \frac{D_k}{N_{2k}}$. The odds ratio then simplifies to $\hat{\alpha}_k = \frac{A_k D_k / T_k}{B_k C_k / T_k}$. When summed over all score categories, the test statistic $\hat{\alpha}_{MH} = \frac{\sum_k A_k D_k / T_k}{\sum_k B_k C_k / T_k}$ becomes a measure of the effective size for the magnitude of DIF. If $\hat{\alpha}_{MH} > 1$ then there is DIF in favor of Group 1, and if $\hat{\alpha}_{MH} < 1$ there is DIF in favor of Group 2. There are several challenges with this scale. While it is intuitive that odds greater than 1 favor Group 1 and odds less than one favor Group 2, the scale is asymmetrical (from 1 to infinity when Group 1 is favored and from 0 to 1 when Group 2 is favored). For ease of interpretability, $\hat{\alpha}_{MH}$ is transformed into a log scale using the formula $\Delta_{MH} = -2.35\, ln\, ln\, (\hat{\alpha}_{MH})$, which is comparable to the Educational Testing Services (ETS) delta scale (Zwick, 2012). On this scale, $\Delta_{MH} = 0$ represents no DIF, negative values indicate that the item favors Group 1, while positive values indicate that the item favors Group 2.

$\hat{\alpha}_{MH}$ *and* Δ_{MH} are both measures of the effect size of DIF. To test for the significance of the DIF statistic, the $M - H_{\chi 2} \sim \chi 2(1)$ and is given by

$$M-H_{\chi^2}=\frac{(|\sum_k A_k-\sum_k E(A_k|-\frac{1}{2})^2}{\sum_k var(A_k)}\frac{2}{}$$ where $E(A_k)=\frac{N_{1k}M_k}{T_k}$ and $var(A_k)=\frac{N_{1k}N_{2k}M_kM_k}{T_k^2(T_k-1)}$

The null hypothesis $H_0: \hat{\alpha}_{MH}=1$ can then be tested for significance in the differences in odds ratios. For large sample sizes, statistical tests will always show significance (Angoff, 1993).

The M-H is a widely used DIF detection method (Bandalos, 2018) by the Educational Testing Services (ETS). Table 4 below presents the ETS criteria for flagging DIF in items. Items displaying DIF are categorized as A, B or C, with A being a negligible DIF, B being a moderate DIF and C being a large DIF (Zwick, 2012).

The M-H makes fewer assumptions about the underlying score distribution and test than other DIF detection methods. It is easy to understand and implement, and the ETS categorization scale provides clear guidelines for the different amounts of DIF (Bandalos, 2018; Zwick, 2012). M-H is said to not need large samples for stable results (Zwick, 2012) compared to IRT-based methods. ETS requires at least 200 examinees for the smaller group and 500 examinees overall when DIF analyses are performed in the test development stage and 300 examinees for the smaller group and 700 examinees overall when DIF analyses are performed in the test scoring phase before scores are released. Effect size and statistical tests help psychometricians avoid being swayed by significance only, even with little effect size.

Identifying an item that displays DIF by either of the methods described in this paper means that this item, which has been proven to function differentially, was used to match examinees based on the assumption that the matching criterion is reliable and valid (in essence, DIF-free). It has been suggested that items which display DIF be excluded from the calculation of the matching criterion (total score) and the DIF analysis be rerun as a way to purify the matching variable (Clauser & Mazor, 1998; Dorans & Holland, 1993; Roussos & Stout, 2004). When this iterative process is applied to the detection of DIF using the M-H in simulation

studies, it has proven to show better results than the non-purified matching criterion (Clauser & Mazor, 1998). However, it is encouraged that the studied item always be included in the total score when DIF is being investigated and removed when other items are studied (Clauser & Mazor, 1998; Dorans & Holland, 1993).

Table 4. ETS Delta Scale

Category	Amount of DIF	Decision Rule
A	Negligible DIF	$\lvert\Delta_{MH}\rvert < 1.0$ and/or $M - H_{\chi^2}$ not significant
B	Intermediate DIF	$1.0 \leq \lvert\Delta_{MH}\rvert < 1.5$ and $M - H_{\chi^2}\ is\ significant$
C	Large DIF	$\lvert\Delta_{MH}\rvert \geq 1.5\ and\ M - H_{\chi^2}\ is\ significant$

Logistic Regression

Logistic regression is used to model the log odds of the probability of getting an item correct as predicted from sum correct scores, *X*, group membership, G, which is dummy coded 0 and 1, and an interaction term of the two. Mathematically, this is represented as

$$ln\left[\frac{p_i}{1-p_i}\right] = \beta_0 + \beta_1 X + \beta_2 \times G + \beta_3 (X \times G)$$

In theory, the coefficient for group membership will be close to 0 and insignificant when there is no DIF. First, just X is entered into the model, G is then added, and finally, the interaction term X×G is added. Log likelihoods (LL) are calculated for each model and

$$Likelihood\ Ration\ (LR) = -2(LL_{smaller} - LL_{larger})$$

is computed for two consecutive models. LR is distributed as a $\chi^2(1)$ and is used to test the null hypothesis that $\beta_i = 0$, $i = (2,3)$. If $\beta_2 \neq 0$, then group membership predicts the log odds of getting the item correct after controlling for X. This is equivalent to the presence of uniform DIF. Similarly, if $\beta_3 \neq 0$, then the interaction of group membership and X is also a predictor of the log odds of getting the item correct indicating the presence of non-uniform DIF.

The inclusion of the total score, X, as a predictor of the log odds of getting an item correct serves the function of matching examinees on ability. It is important to note that X is treated as a continuous variable and excludes the score on the examined item.

Logistic regression is a robust, practical, simple-to-use method for DIF detection (Healy, 2006) that can be used with sample sizes as small as 200 in each group (Camilli & Shepard, 1994, as cited in Bandalos, 2018). It has provision for including other predictors such as other abilities (Clauser & Mazor, 1998). Logistic regression tests for uniform and non-uniform DIF (Bandalos, 2018). In logistic regression, the total score is treated as continuous as opposed to M-H, which needs to include a continuity correction (Clauser & Mazor, 1998).

IRT Methods

Unidimensional item response theory (IRT) models the probability of success on an item as a function of a latent trait (ability). In the 3PL model, the probability of getting an item correct given ability of θ is:

$$P(Y_j = 1 \mid \theta) = c_j + (1 - c_j)\,\frac{exp[a_j(\theta - b_j)]}{1 + exp\ exp\ [a_j(\theta - b_j)]}$$

where a_j is the item discrimination parameter, b_j is the item difficulty parameter, and c_j is a pseudo-guessing parameter. The 2PL model is obtained by setting $c_j = 0$ and the 1PL model by further setting $a_j = 1$. Response data is used to estimate the item parameters (calibration), and they are hypothesized to be stable over administrations and examinees. All the IRT approaches

to detecting DIF discussed in this paper involve calibrating the model separately for the two manifest groups and calibrating the items with all examinees as an intact group. Three IRT approaches are covered in this paper: likelihood ratio tests, which compare model fit with and without separate calibration; parameter comparison tests, which test for the difference in item parameters for the two manifest groups; and the area-between method, which quantifies and tests the area between ICCs for the two groups.

In general, IRT methods need a large number of examinees in each of the groups for stable parameter estimation, particularly if the 2PL (>500 examinees) and 3PL (>1,000 examinees) models are used (Clauser & Mazor, 1998).

Thissen-Steinberg-Wainer (TSW) Likelihood Ratio Test

As stated earlier, IRT-based methods for DIF detection include the separate calibration of item parameters. In this method, the IRT model of choice (1PL, 2PL or 3PL) is calibrated for all examinees with the parameters constrained to be equal (C) and then calibrated with the parameters allowed to vary for Group 1 and Group 2 for the item under investigation (A) (Bandalos, 2018; de Ayala, 2009; Osterlind & Everson, 2009). The likelihood ratio is given by $G^2 = -2\ ln\ \frac{L(C)}{L(A)} \sim \chi^2_{df}$ where df = the number of parameters allowed to vary.

If an item does not display DIF, there is no significant difference in the estimated item parameters, and both calibrations should fit the data equally well. If, however, there is a significant difference in the model fit, then the item displays DIF.

Different approaches to detecting DIF using likelihood ratio tests exist, mainly differing on the number of parameters allowed to vary and the order in which they are allowed to vary. A significant G^2 statistic implies that the parameter(s) allowed to vary are different for the two groups for that item (Bandalos, 2018; Osterlind & Everson, 2009).

Lord's Parameter Comparison

Another approach to the detection of DIF within the framework of IRT is to test for the difference in estimated item parameters when separate calibrations are performed for each manifest group. The indeterminacy of the ability the scale means that scale for ability (θ) is unique to each group. In order to compare item parameters, linking is undertaken to transform the estimates for Group 2 into the metric for Group 1 (de Ayala, 2009; Shultz et al., 2014).

The Wald statistic below is used to test for the difference in maximum likelihood estimators (MLE) of the difficulty parameters for the two groups (Thissen et al., 1993). If only the difference in difficulty parameters is considered, then the Wald statistic for items i and d_i is given by

$$d_i^2 = \left[\frac{\hat{b}_{Group\ 1} - \hat{b}_{Group\ 2}}{SE(\hat{b}_{Group\ 1} - \hat{b}_{Group\ 2})}\right]^2 \sim \chi_1^2$$

For the 3PL model, discrimination, difficulty, and pseudo-guessing parameters are compared. The more general Wald statistic is given by

$D^2 = \hat{v}^T \hat{\Sigma}^{-1} \hat{v} \sim \chi_3^2$ where $\hat{v} = \left(\frac{\hat{a}_{Group\ 1} - \hat{a}_{Group\ 2}}{\hat{b}_{Group\ 1} - \hat{b}_{Group\ 2}}\ \hat{c}_{Group\ 1} - \hat{c}_{Group\ 2}\right)$ and $\hat{\Sigma}$ is the estimate for the sample variance-covariance matrix for the differences between the parameters.

$D^2 \sim \chi^2(2)$ and can be used to test the null hypothesis H_0: $D^2 = 0$. If the difference in item parameters for Group 1 and Group 2 is significant then the item displays DIF.

de Ayala (2009) outlines further assumptions of IRT in general and hence extend to the use of IRT methods for DIF detection:

- Local independence.

  Local independence refers to the independence of responses on two different items. An examinees response on an item depends only on their

ability (θ) and not on how they respond to any other item on the test. This assumption would be violated if the test contains items based on a common text for instance. Mathematically, this can be expressed as

$$P(Y_j = 1, Y_{j'} = 1 \mid \theta) = P(Y_j = 1 \mid \theta)P(Y_{j'} = 1 \mid \theta) \quad (j \neq j')$$

- Good model-data fit.

IRT-based DIF detection methods assume that the data fit the function specified by the model used (1PL, 2PL, or 3PL). This is particularly important when the Likelihood Ratio Function is used for DIF detection because it relies on the differences in fit of the models when item parameters for the two groups are calibrated concurrently and separately.

In general, IRT methods are considered more comprehensive (when the 3PL model is used) because they test for differences in all three parameters of the IRT model (Angoff, 1993). When the 2PL model is used, that is, when guessing is ignored or when the guessing parameter is assumed, and/or constrained to be equal in both groups, IRT-based methods can detect both uniform and non-uniform DIF.

Likelihood ratio tests and the parameter comparison method can indicate effective sizes, the amount of DIF displayed, and statistical tests of significance, whereas area measures only indicate how much DIF is present but not whether it is statistically significant.

Multidimensional Approaches

The primary occupation of DIF detection procedures that fall into the categories above has been to flag items with DIF as potentially biased. Attempts to determine the causes of DIF post hoc through substantive reviews are reported to be largely unsuccessful (Roussos & Stout, 1996). Often, these substantive investigations include enlisting content matter experts to identify the cause of the DIF and provide a rational reason why the focal group has a different correct

response probability. Both IRT and non-IRT-based approaches approach the analysis of bias by separating the statistical detection of DIF from the substantive investigation of the causes of DIF. In these approaches, DIF is first detected and depending on the magnitude detected, substantive analyses such as content expert review are conducted to determine the cause of DIF.

Another critique of these approaches to DIF detection is that they are conducted at the item level with no formal approach to examining how DIF items interact or compound at the test level (Zumbo, 2007).

The multidimensionality-based DIF approach reverses the order of bias investigations by performing the substantive, theory-generating investigations in the first stage and collecting data to confirm or disconfirm the hypotheses (Gierl, 2005; Shealy & Stout, 1993). This approach, in addition to testing the multidimensionality-based DIF hypothesis, is also concerned with circling results back to the test development phase to aid the design of DIF-free items (Ackerman et al., 2003; Roussos & Stout, 1996). This theoretical focus on why DIF occurs marks what Zumbo (2007) termed the third generation of DIF.

Others, such as Benitez et al (2016), have leveraged the mixed methods and integrated quantitative DIF analyses with independently solicited qualitative responses from bilingual experts on possible reasons for the incomparability of responses from translated versions of the same item. Similarly, in their proposal for the integration of DIF analyses to all five sources of valid evidence, giving DIF analyses a more central role in the validation of tests/measures, (Gómez-Benito et al., 2018) also argue that the mixed methods research framework is most appropriate.

Race and Ethnicity

Race is "the grouping or classification of people based on what are presumed to be biological differences typically evident as differences in physical differences due to such features as skin color" (Kivisto & Croll, 2012). Conceptions of race solidified along color lines in tandem with European colonization (Jackson & Weidman, 2004; Kivisto & Croll, 2012; Zuberi, 2001). Ethnicity is reported to have come into use in the 1930s in tandem with the rise of Nazi Germany and refers to a "collective identity based on a subjective belief in a shared culture" (Kivisto & Croll, 2012).

Despite progressive proclamations on the socially constructed nature of race, conceptions of race as a fixed attribute persist (Poe, 2009; Ross et al., 2020; Smedley & Smedley, 2005; Zuberi, 2001). In particular, "when race is treated as a fixed characteristic, it becomes a variable from which causal explanations can be assigned" (Zuberi, 2000), which has perpetuated the pathologizing of certain races in favor of the White race (Valencia, 1997). In program evaluation as far back as the 90s, take a critical look at how differing conceptions of race bear on evaluation work and how findings are used. Biogenetic conceptions of race, when used without the requisite skills in those fields, lead to myopic and potentially harmful uses of evaluation findings (Davis, 1992).

As discussed above, one of the central concepts of CRT is the pervasive nature of race despite its socially constructed nature. Smedley & Smedley (2005) and Torres & Collon (2015) posit that conceptions of race as a social construct are restricted as it does not hone in on the experience of racialization. Critical quantitative researchers underscore that race is an important variable but cannot be used uncritically. When race is conceptualized as a socio-cultural, it needs to be considered in the concomitant sociocultural context. One theme in the literature of race and

ethnicity as variables in research is the assertion that they act as proxies for a plethora of social variables such as socioeconomic status, parental education level, access to resources and so on. Thus, "without defining race-ethnicity and commenting on how social processes might result in research findings (i.e., how race is constructed in a specific context), researchers run the risk of reinforcing notions of the fixed-ness of racial-ethnic identity." (Ross et al., 2020).

Roth (2016) condensed the multiple dimensions of race measured in research into a typology which she links to outcomes that are appropriate for each dimension. She differentiates racial self-identification which allows participants to subjectively state their race from racial self-classification which is often used on official forms and requires participants to choose from a prescribed set of racial identities. Her typology goes on to describe observed race, which can either be based on one's appearance or one's interactions, which is often measured by enumerators or interviewers and is the race others believe one to be. It also highlights race as ancestry which is either linked to the racial groups of one's known ancestors or increasingly in health research, genetic ancestry testing.

The use of race as a variable in research has been explored more in health and medical research (e.g. Manly, 2006; Mateos et al., 2009; Ross et al., 2020; Torres & Colón, 2015), psychology (e.g. Smedley & Smedley, 2005; Zuckerman, 1990), and program evaluation (e.g. Davis, 1992). For instance (Baker, Dominique et al., 2022), in their systematic literature review of AERA-published articles between 2009 and 2019, explored how researchers describe the racial categories in their articles and found that there was a wide variation in the inclusion of racial groups in analyses between journals and even within the same journal, over time and at any given point.

In education, evaluation and health research, definitions and theoretical framing of the race and ethnicity variables still need to be improved. Another common criticism of the use of race as a variable in research is that for various reasons, including small sample sizes, some racial-ethnic groups are excluded from analysis without explanation (Ross et al., 2020). In addition, the dichotomizing of racial analyses into Black or White leaves other minoritized groups (Castagno, 2005).

Race and Ethnicity in DIF

The impetus for DIF studies investigating bias in test items after the Civil Rights Act of 1964 has been stated previously. Despite this centrality of race and ethnicity in analyses such as DIF, the field of measurement and psychometrics has produced little research on the conceptualization, operationalization or interpretation of race and ethnicity as variables, causal or otherwise, in psychometric research. A search for "race", "ethnicity", or "minorities" in the top 9 psychometric journals[1] from 1940 to 2022 yielded between 4 and 78 articles each, but none related to the definition or operationalization of race or ethnicity as variables in psychometric analyses. In the same vein, introductory psychometric textbooks (e.g. Bandalos, 2018; Shultz et al., 2014) also do not define race or ethnicity nor clarify how they are operationalized as variables in the conduct of measurement procedures.

[1] Psychological Methods (APA); Psychometrika (Psychometric Society); Journal of Educational and Behavior Statistics (AERA and American Statistical Association); British Journal of Mathematical and Statistical Psychology (British Psychological Society); Multivariate Behavioral Research (Society of Multivariate Experimental Psychology); Applied Psychological Measurement (SAGE Publishing), Journal of Educational Measurement (NCME), Educational and Psychological Measurement (SAGE Publishing), Educational Measurement: Issues and Practice (NCME)

The Standards

The most recent 2014 Standards for Educational and Psychological Testing has 26 references to race and ethnicity, with Chapter 3 on Fairness in Testing having the highest number of such references (n = 9). In all references, race and ethnicity are referred to in conjunction with other variables such as gender, age, socioeconomic status, language background, and so on that partition the population of examinees.

Throughout the Standards, race and ethnicity are presented in two different though related ways. The first is as variables that partition the examinee population (*group*) into relevant *subgroups*. A "subgroup includes members of the larger group who are identifiable in some way that is relevant to the standard being applied" (p. 6). For example, "… groups defined in terms of race/ethnicity, gender, age, and other characteristics" (p. 20) or "… diverse subgroups such as those defined by race, ethnicity, gender, culture, language, age, disability or socioeconomic status" (p. 49). As the Standards are not intended to be followed in a cookbook or checklist fashion, professional judgement is required in the determination of which subgroups are to be considered, though in some cases, legal requirements may mandate which subgroups are relevant (American Educational Research Association et al., 2014, p. 6).

The second way race and ethnicity are used in the Standards is as demographic characteristics that take on predetermined values by each member of the population. In this instance, race and ethnicity are among other characteristics, such as socioeconomic status. In Chapter 7, the Standards outline the supporting documentation for tests and demographic characteristics, including race and ethnicity, are recommended to be collected for the groups and individuals who participate in test development and validation studies.

On four occasions, race and ethnicity are posited as factors that may affect test performance. For instance, in Chapter 3, race and ethnicity are listed among other "factors that may affect the performance of the test taker" along with other demographic variables such as gender, and linguistic and cultural background of both examiner and test taker and also situational variables such as "test taker's experience with formal education, the testing style of the examiner, the level of acculturation of the test taker and examiner, the test taker's primary language, the language used for test administration (if it is not the primary language of the test taker), and the use of a bilingual or bicultural interpreter." (p. 55). In Chapter 10 on Psychological Testing and Assessment, race and ethnicity are listed as two of the many factors that "may influence individual test results and the overall outcome of the psychological assessment." (p. 167). To this extent, the Standard recommends that test scores be considered in relation to other qualitative observations.

Whether considered as a demographic characteristic or as a determinant of group membership, The Standards at times present race and ethnicity as determinants or causes of performance or lack thereof on tests as presented in test scores. Moreover, the broad scope of application of the Standards is a good reason to keep definitions inexplicit to facilitate the applicability of Standards to different contexts. However, the description of the lack of definitions for race and ethnicity and the theoretical basis for their conceptualization as variables that affect performance or test scores is missing in the Standards.

Conclusion

In this chapter I have introduced the critical quantitative elements that framed this study. I briefly introduced the history of psychometrics and testing and bias in testing. I outlined DIF and described the most common IRT and non-IRT DIF detection methods. I outlined

multidimensional approaches to DIF detection. I provided an overview of race and ethnicity as variables in research. The chapter concluded with a critical review of how race and ethnicity are under conceptualized in The Standards for Educational and Psychological Testing. Despite the use of race and ethnicity as grouping variables, there was no literature on their conceptualization in the field of psychometrics. The next chapter describes the methodology for this study.

CHAPTER III: METHODOLOGY

This chapter presents the research design for this is study that addressed the research questions presented in Chapter II. It begins with a rearticulation of the research questions, proceed to provide an overview and description of the sequential mixed methods design that was be used and proceeds to explain each phase in detail.

Research Questions

RQ1: What are key characteristics of the differential item functioning investigations that employ race and/or ethnicity as grouping variables and appear in peer reviewed journal articles published 2015-2020 literature?

RQ2: What trends emerge in differential item functioning (DIF) analyses reported in recently published research studies that use race and/or ethnicity as grouping variables in terms of

a) how the terms are defined and operationalized,

b) how the categorization is conducted, and

c) how findings are reported and interpreted?

RQ3: How does the conceptualization of race and ethnicity in recently published DIF studies by the researchers identified for RQ1 respond to the particularities of the study contexts?

RQ4: To what extent do findings from DIF studies that use race and/or ethnicity as grouping variables align with the authors' they cite in terms of interpretation when utilized in future research?

Research Design

A sequential explanatory mixed methods design (Creswell & Plano Clark, 2018), illustrated in The interpretation of DIF findings had four categories: no interpretation of DIF findings, presentation of DIF findings with no connection to the grouping variable (casual interpretations), an interpretation of DIF findings that is connected to the grouping variables and interpretation that not only connects findings to the grouping variable but goes further to connect DIF findings to the broader socio-political context. This variable was used to stratify the sample of research studies in preparation for the quantitative phase.

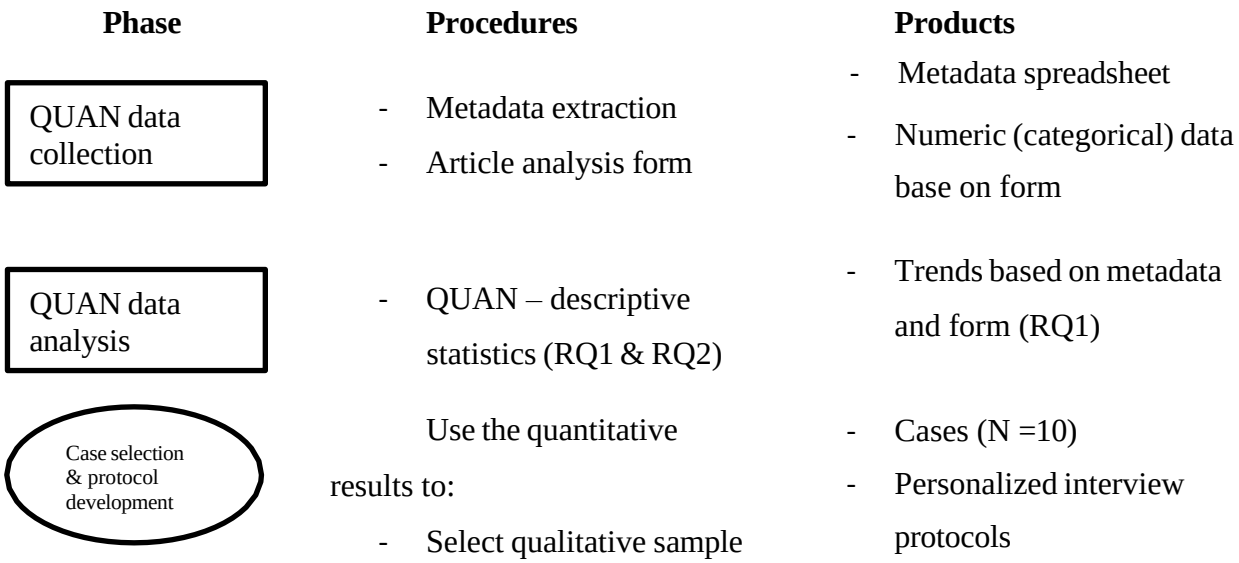
In the qualitative strand of the study, 30 researchers from the last two categories identified in the case selection phase were invited for interviews to explore in greater depth their conceptualization of the race and ethnicity variables, their rationale and procedures when using the variable for grouping in DIF analyses and their interpretation of the findings. In addition, two more studies were included in Phase 2 based on the explicit definition of race and/or ethnicity provided in the manuscript. A total of 32 US-based researchers were invited to reflect on how their work is being cited, with the aid of the forward citations (publications made after an article is published that include that article in their reference list) gathered in the case selection phase.

Figure 2 below, will be used in this study to leverage the benefits of both quantitative and qualitative data collection and analysis. The quantitative phase of the study will investigate the trends in recently published studies of differential item functioning analyses between 2015 and 2020 through a systematic review. The importance of this phase is to gain a broad view of how the race and ethnicity variables are conceptualized and used in contemporary studies of DIF and how interpretations of DIF findings are reported concerning the grouping variable(s).

The interpretation of DIF findings had four categories: no interpretation of DIF findings, presentation of DIF findings with no connection to the grouping variable (casual interpretations), an interpretation of DIF findings that is connected to the grouping variables and interpretation that not only connects findings to the grouping variable but goes further to connect DIF findings to the broader socio-political context. This variable was used to stratify the sample of research studies in preparation for the quantitative phase.

In the qualitative strand of the study, 30 researchers from the last two categories identified in the case selection phase were invited for interviews to explore in greater depth their conceptualization of the race and ethnicity variables, their rationale and procedures when using the variable for grouping in DIF analyses and their interpretation of the findings. In addition, two more studies were included in Phase 2 based on the explicit definition of race and/or ethnicity provided in the manuscript. A total of 32 US-based researchers were invited to reflect on how their work is being cited, with the aid of the forward citations (publications made after an article is published that include that article in their reference list) gathered in the case selection phase.

Figure 2. Mixed Methods Design Diagram



	- Develop a personalized interview protocol	
QUAL data collection	- Interviews - Narrative of the articles - Field notes	- Text data (interview transcripts, documents)
QUAL data analysis	- Deductive coding within - Inductive coding across (RQ3 & RQ4)	- Themes
Integration of QUAN and QUAL results	- Summarize and interpret QUAN results - Summarize and interpret the QUAL results - Discuss to what extent and in what ways the qualitative results help to explain the quantitative results	- Discussion - Implications - Future research

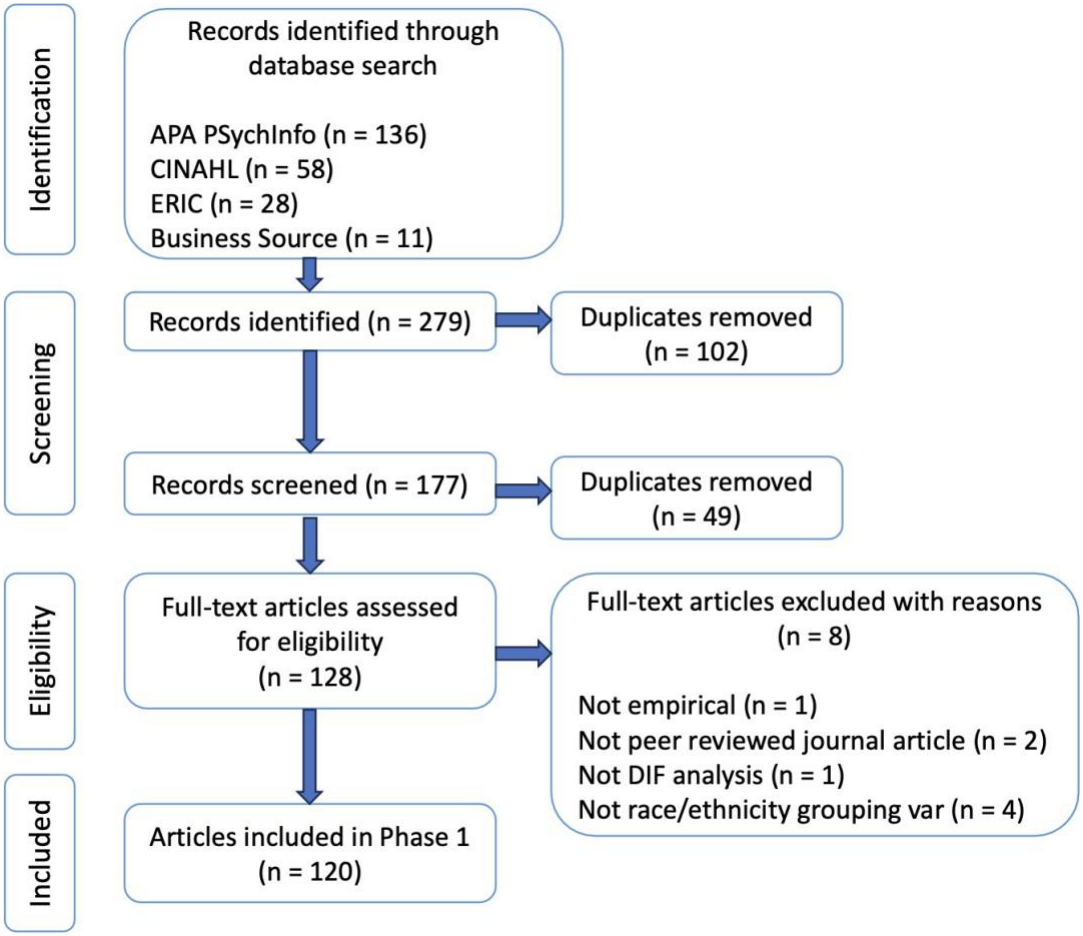
Sources of Data

Research articles published in peer-reviewed journals were the first source of data for this study. To align with other systematic reviews related to DIF analyses, ERIC and PsychInfo will be the databases used to search for articles (Gomez et al., 2005; Berrio et al., 2020). Articles with the key phrase “differential item functioning” or “DIF” and either “race” or “ethnicity” appearing in the abstract were considered for inclusion. In addition, articles should include “empirical study” as the methodology. Five academic databases, APA PsychInfo, ERIC,

Education Source, Business Source and CINAHL, were searched, and all articles published between 2015 and 2020 that met the search criteria were considered for inclusion. To accommodate the second phase of the study, it was important to set a cutoff date that would allow sufficient time for the core DIF article to be cited.

Figure 3 below is the PRISMA diagram the summarizes the systematic data searc. 279 records were identified from the four academic databases: APA PsychInfo (n=136), CINAHL (n=58), ERIC (n-28) and Business Source (n=11). After removing duplicates 177 record were screened and a further 49 duplicates were removed. 128 articles were assessed for eligibility resulting in the exclusion of 8 articles for reasons such as not being empirical, not from a peer-reviewed, not including DIF analysis or not focusing on race or ethnicity as grouping variables. Ultimately 120 articles were included in Phase 1 of the study.

Figure 3. PRISMA Diagram for Literature Search



Phase 1 – Quantitative Phase

Data Collection

As part of data cleaning, duplicate studies will be excluded from the list of identified studies. For articles that are indexed but not publicly available, the corresponding author will be contacted via email to request access to their study manuscript. If there is no response, two follow-up emails will be sent, following which the article will be reported as irretrievable. Articles that meet the criteria will be collected, and their abstracts and/or methods sections will be screened to ensure that they include an empirical study (i.e., not a theoretical article or a simulation study), that DIF is included in the analysis and that race and/or ethnicity are used as a grouping variable. In addition, articles that apply or demonstrate new DIF detection methods or simulation studies were excluded.

To ensure the trustworthiness of the data collected, I enlisted the assistance of a peer to independently implement my inclusion/exclusion criteria for 20% of my identified research studies. The inclusion criteria were 1) an empirical study that 2) includes at least one DIF analysis with 3) race and/or ethnicity used as a grouping variable for the DIF analysis. Articles were excluded if they 1) used simulated data or 2) were an application or demonstration of a new method. The independent rater was also presented with abstracts for 20% of the 177 abstracts identified for screening. Interrater reliability was calculated by comparing the proportion of inclusion/exclusion decisions that agreed. An interrater reliability of 94% was achieved. Disagreement occurred for 2 articles where I had marked as “Not sure” while the independent reviewer marked “Exclude”. All articles marked “Not sure” were retrieved, and the methodology section was further reviewed to determine if they met the inclusion criteria. After the second review, the agreement was 100%.

Quantitative data was generated from identified research articles published in peer-reviewed journals using a quantification form (Table 5). In addition, to study metadata, study details relating to the DIF analyses will be recorded, including the type of test studied, details of the population the test was administered to, the type of DIF detection method used, the name of the dataset (if it was not primary data collected by the researcher(s).

Table 5. Data Quantification Form

Question	**Rules for responses**
Article ID	**ID** (assigned to articles prior to coding)
Title	**Free response**
Author(s)	**Free response**
Field	**Education** **Psychology** **Public health** **General research** **Other =** Free response
Type of study	**Quantitative** **Mixed Methods**
Primary or secondary data	**Primary** **Secondary**
Name of dataset (if secondary data is used)	**Free response**
Type of test studied	**Proficiency** **Achievement** **Aptitude** **Intelligence** **Personality** **Placement** **Licensure/certification** **Diagnostic** **Other** = Free response
Examinee population age range	**Free response** **Not reported**

Age of respondents	**0 – 5** **K – 12** **College** **Adults** **Seniors** **Other** = Free response
Total number of examinees	**Free response**
Study conceptualization	**Fairness and equity in testing** **Investigating threats to internal validity** **Comparability of translated/adapted tests** **Understanding item response processes** **Investigating measurement invariance** **Validation of a new instrument** **Other** = Free response
Theoretical framing for DIF	**Free response** (paste section and provide page number)
Race/ethnicity grouping variable	**Race** **Ethnicity** **Race/ethnicity** (Select all that apply)
Justification for use of the race/ethnicity variable	**Free response**
Race/ethnicity categories used	**White/Caucasian** **Black/African American/Afro American** **Asian/Asian American** **Native Hawaiian/Pacific Islander** **Native American/American Indian/Alaska(n) Native** **Two of more mixed race/Multiracial/Multiethnic**
Race/ethnicity categories excluded	**White/Caucasian** **Black/African American/Afro American** **Asian/Asian American** **Native Hawaiian/Pacific Islander** **Native American/American Indian/Alaska(n) Native** **Two of more mixed race/Multiracial/Multiethnic**
How are participants allocated into race/ethnicity categories	**Researcher assigned** **Participant self-selected** **Third party records (medical, school, census, etc.)** **Other** = Free response
Other grouping variables used	**Gender** **Age** **Language background**

	Language proficiency **Immigration status** **Health status** **Other** = Free response
Total number of items	**Free response**
DIF Detection Method	**Free response**
# items flagged for DIF	**Free response**
Proportion of items flagged for DIF	**Computed variable (Number of items flagged for DIF / Total number of items)**
Interpretation of findings with respect to racial/ethnicity groups	**No interpretation - no interpretation** **Some interpretation – DIF results mentioned but not interpreted** **Full interpretation – findings linked to race/ethnicity groups** **Contextualized interpretation**
Notes/Observations	**Free response**

The study conceptualization was coded according to Zumbo's (2007) purposes for DIF studies, as detailed in Table 6. Study Conceptualization below. Studies will be categorized as 'Fairness and equity in testing' when the purpose of the DIF and the impetus for the investigation is the protection of visible minorities or language groups as mandated by legislation or policies. Studies will be categorized as 'Investigating threats to internal validity' when DIF needs to be ruled out to make comparisons of the performance between groups. Similarly, but in international or cross-cultural contexts, 'Comparability of translated/adapted tests' will be used when tests are adapted and translated to ensure that differences in the performance of groups are not a result of DIF. The categorization 'Understanding response processes' will be used for studies that frame DIF studies to understand test takers' response processes. 'Investigating lack of invariance' will include studies that conduct DIF analysis to ensure the suitability of the IRT models used. These studies often include establishing a lack of measurement invariance and

assessing model-data fit. An additional category, ‘Validation of a new test/measure’, will be used as a slight variation on the previous category to identify studies conducted when a new instrument/test/measure is being created. Such studies, in addition to investigating measurement invariance and model-data fit, often include selecting optimum items from an item bank wherein items that display DIF are discarded or not considered for inclusion.

Specifically related to race and ethnicity as grouping variables, I recorded the rationale for using the variables (when it was provided), how participants/examinees were allocated into their respective racial/ethnic groups (researcher assigned, participant self-selected or collected from a third party such as health or school records), how many racial and ethnic groups were considered (including the number of participants in each group), which populations (if any) were reported to be excluded or collapsed. For each study, other grouping variables used, such as gender, language background, education level, and so on, were recorded.

Table 6. Study Conceptualization

Category	Description	Examples
Investigating/establishing measurement invariance	Empirical method for investigating: (a) lack of invariance, (b) model-data fit, and (c) model appropriateness Precursor to model-based statistical measurement frameworks	““Thus, the goal of this set of analyses was to use latent variable modeling in the development of a short-form version of the MMT. Latent variable modeling, including factor analyses and IRT, was used to examine the psychometric properties of the MMT and the resulting short-form, including dimensionality, reliability, information and measurement equivalence. This is the first application of IRT to examine a medication management test.” *(Teresi et al., 2018)* “In this study, we apply the Rasch measurement model to further evaluate the POM. Specifically, we investigated person and item fit statistics, response scale, dimensionality of the scale, and differential item functioning.” *(Cordier et al., 2019)*
Investigating threats to internal validity	Precursor to making group comparisons. Groups identified ahead of time.	“Measurement equivalence of the ICH-CAHPS survey is needed for comparisons of patient experiences from different ICH subgroups,

	Decided in line with RQs Developed as DIF moved into day-to-day research settings	especially if they are used to assess disparities associated with gender, age, race, and education." *(Setodji et al., 2019)* "DIF analysis should be performed before comparing composite scores across subgroups of other sampled populations." *(Janulis et al., 2018)*
Comparability of translated / adapted tests	International, comparative, and cross-cultural research Comparing translated tests Comparing adapted tests	"However, this is the first study to investigate cross-linguistic and cross-cultural validity of the Turkish version of the PHQ-9, and one of the few to study this topic at all. Consequently, all items of the PHQ-9 were tested on DIF without statistical pre-assumptions. Based on the results, recommendations for applying the PHQ-9 in Turkish immigrants are provided." *(Reich et al., 2018)*
Fairness & equity in testing	Tied to policy and legislation. Groups defined ahead of time.	"The goals of this study were to (1) evaluate the degree to which the ECog (Everyday Cognition) provides an unbiased assessment of functional abilities across three ethnoracial (ER) groups of older adults [non-Hispanic White (NHW), Hispanic, and Black]" *(Filshtein et al., 2020)*

Understanding item response processes	Understand the cognitive and/or psychosocial processes. Are these processes the same for different groups of individuals? Considering bounds and limitations of measurement inferences. Groups identified ahead of time OR Groups identified by latent classes.	"There is still a need for additional research to address the theory that distinct cultural experiences may differentially affect the formulation of racial/ ethnic identity, which may in turn influence how the racial/ethnic identity construct functions across different geographical contexts (Cokley, 2007)." *Chakawa (2015)* "Given the large number of bilingual Spanish-English speaking students in the United States and the variety of factors that can contribute to language proficiency, it is important to understand the way in which items on standardized tests function for diverse Spanish-English bilinguals." *(Sandilos et al., 2015)*
Validation of new instrument	Part of validation studies in the development of a new measure	"The purpose of this study was to provide a psychometric assessment of the DBS (Disclosure Belief Scale) instrument for validation purpose and to facilitate future research in assessing individual's serostatus disclosure belief." (Hu et al., 2017)

Regarding the findings of the DIF studies, the form recorded the number and proportion of items displaying DIF. The extent of the interpretation of DIF results was recorded as '**No interpretation**', where there was at least one item with detectable DIF. However, DIF results were presented in results tables and re-stated in the narrative. Articles were categorized as providing '**Some interpretation**' if DIF findings were presented in tables and figures and described in the narrative of the results section but not interpreted in relation to the race/ethnicity grouping variable. The '**Full interpretation**' category referred to instances where DIF findings were presented in tables, in the narrative, and linked to the grouping variable. The direction of DIF was provided, while articles were categorized as '**Contextualized interpretation** when they provided interpretation of DIF findings that were presented in tables and narrative. This can be linked back to the race and/or ethnicity grouping variable(s) and situated the findings in the broader sociopolitical context of the respective study

Data Analysis

The unit of analysis in the quantitative phase of the research was each research article. Since all the numeric data generated in this phase were categorical, the analysis was descriptive. Data in frequencies and percentages summarized the trends and tendencies in published studies of DIF. In addition, a cross-tabulation of the interpretation of findings for each of the different conceptualizations of the study will be compiled to gain insight into how different study conceptualizations interact with the interpretation of DIF results.

Case Selection

After data analysis was complete for Phase 1, articles based in the US were identified as providing a full or fully contextualized interpretation of DIF findings (N = 36). In addition, two articles that did not provide a full or contextualized interpretation of DIF findings but defined

race and/or ethnicity within the manuscript. Email invitations were sent to the first or corresponding authors of the 38 articles identified in November 2023, and reminders were sent two weeks later. Two authors responded to the invitations and were interviewed as part of Phase 2.

To facilitate a generative discussion with researchers in Phase 2 of the study, forward citations will be gathered using Google Scholar. A numbered list of all studies that cite the article will be compiled. All accessible articles on each list will be retrieved, and data reduction will consist of extracting the paragraph(s) that cite the DIF study. The reference list and excerpts will be compiled in a document that will be included in the recruitment email and form part of the interview protocol. At this stage, any additional interview questions specific to the research article or category of a research article will be included in the interview protocol.

Phase 2 – Qualitative Phase

Data Collection

In the qualitative strand of the study, interviews with researchers will be conducted virtually or in person where possible. With their permission, interviews will be audio-recorded and transcribed. A summary document and the transcript will be sent to interviewees to check for accuracy. Once transcriptions are cleaned and the member checking is completed, the audio recordings will be destroyed. Transcripts will be stored in a university-approved secure cloud storage location and retained for 5 years, after which they will be permanently destroyed.

The interview protocol used in the qualitative phase will invite authors to discuss their rationale for using DIF as part of their study, the procedural details of the analyses and their interpretation of the findings. The protocol will also allow researchers to see through the forward citations how their work is being cited and/or used and reflected. The protocol invites researchers

to think about what they would do differently if they were to write up the research findings again, knowing what they know about how the findings are used. This is important because research findings from studies that use, critically or otherwise, race or ethnicity as grouping variables can lend themselves to use or misuse. It also invites them to reflect on any additional considerations they might make for future studies that include DIF. This will hopefully provide consideration to other researchers who use race and ethnicity as grouping variables in DIF studies.

Data Analysis

The analysis of interview data will be analyzed in 5 steps, as depicted in Figure 4 below. The first three steps will be conducted independently within each transcript, while the last two steps will be conducted across transcripts. The unit of analysis will be each interaction (response to an interview question and clarifying questions).

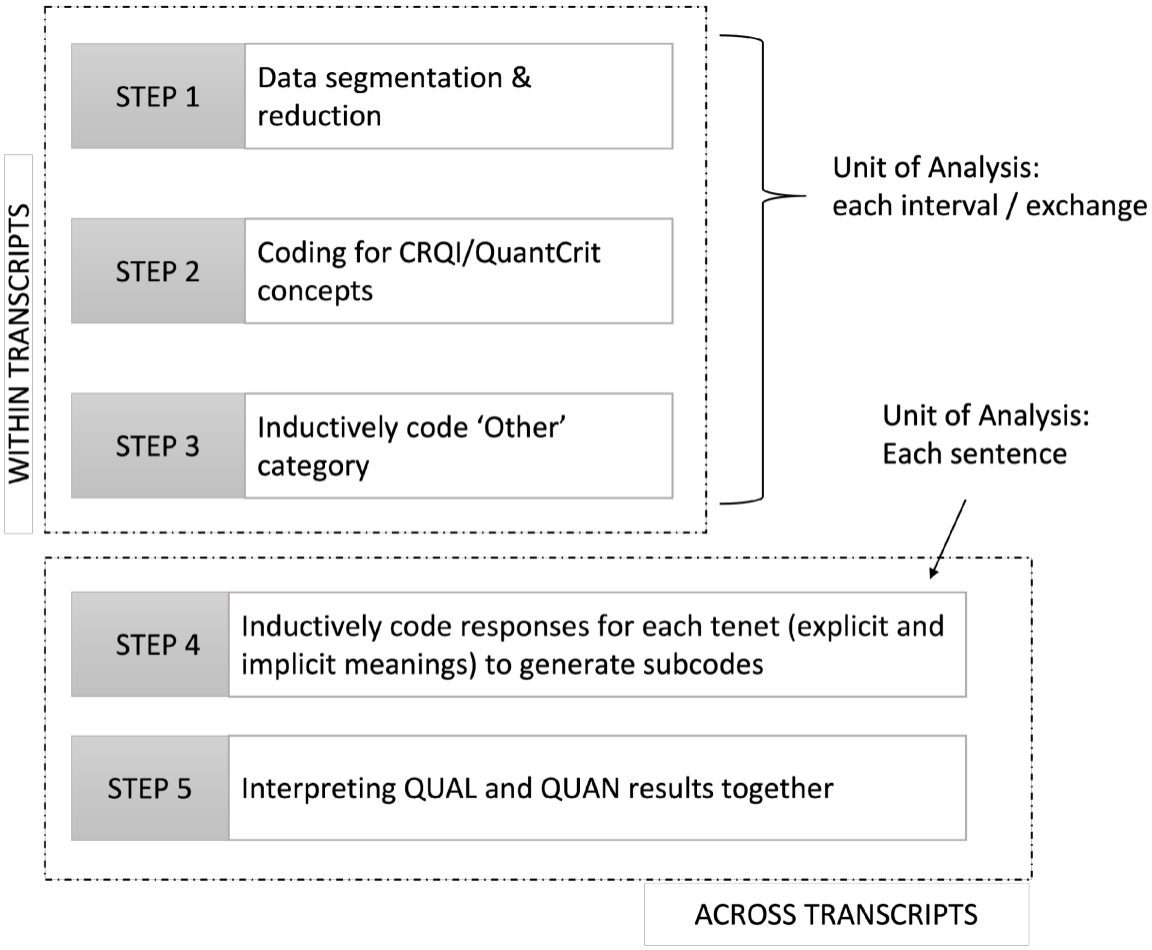
In step 1, the responses/interactions will be segmented according to which research question they address. Data reduction will occur at this stage as I remove interactions unrelated to the study. In step 2, each interaction will be deductively coded for the QuantCrit and CRQI tenets it addresses. At this level, as far as possible, there will be no double coding of responses. A category labelled ‘other’ will be applied to responses that do not fall neatly into the framework. Step 3 will be line-by-line coding of the ‘other’ category to explore whether additional concepts are required or if inductive codes in a pseudo-grounded theory approach would be better suited.

Once these steps were carried out for each interview transcript, step 4 looked across the transcripts for each concept to inductively develop the subcodes for each tenet. This phase will

substantiate the tenets by providing insight into what researchers say in relation to that particular tenet or concept.

In the fifth and final stage, the qualitative findings will be displayed and discussed with the findings of Phase 1 in a joint display. They will better understand how the race and/or ethnicity variables are conceptualized and used in DIF studies.

Figure 4. Qualitative Data Analysis Map



Quality Assurance

In this study's quantitative phase, I intend to collect all publicly available peer-reviewed studies that include DIF analyses and use race and/or ethnicity as grouping variables. I will attempt to ensure data quality when searching by articulating the inclusion and exclusion criteria.

I will enlist the help of a colleague to independently code (include/exclude) 10-15% of the research studies found from the literature search and compare their include/exclude decisions against my own. If the agreement on study inclusion is below 90%, we will discuss the studies and reasons for the discrepancy and refine the inclusion-exclusion criteria until an agreement of 95% is achieved. In the extraction of variables from each study, I will produce a codebook with operational definitions and examples (and counterexamples) for each variable category. At this stage, a colleague will also be enlisted to independently code 10-15% of the studies identified, and the codebook will be refined until an agreement of 95% is achieved.

In the qualitative phase of the study, my role as a researcher will be more pronounced. In preparation for this phase, forward citations of the DIF studies selected will be collected and collated for reflection by the researcher. I intend to produce a summary of the interviews with each participant that will be sent to them within 7 days for review and member checking. This is intended to provide each researcher with highlights from our conversation and an opportunity to clarify any aspects of the responses they wish. In addition, I intend to keep a research journal where I will make entries before and after each interview. I will go back to this journal often to review entries from previous interviews and ensure I can become aware of and confront my research biases. These entries will also be considered when coding interview transcripts. In the write-up of research findings from the qualitative phase, thick and rich descriptions will be provided.

Ethical Considerations

In compliance with the Institutional Review Board (IRB) regulations, an application will be submitted and approved. Data collection tools for both phases of the study have been developed and will form part of the IRB approval submission. A recruitment email introducing

the study and explaining how the researchers were selected for inclusion was also developed and is appended to this proposal (Appendix B). The recruitment email included a full list of forward citations and a selection of excerpts. An informed consent form has also been developed to guarantee participants' rights if they agree to participate in the study (Appendix C).

Participant's identities were protected to the extent possible by:
(1) using pseudo names on all transcripts and storing the master file with author names, studies, and pseudo names in a separate document that is not shared with anyone other than myself,
(2) destroying audio/video recordings after transcripts have been generated and cleaned,
(3) storing all study data in a university-approved, password-protected cloud storage folder with access limited only to myself and the advisory committee.

Positionality Statement

I begin this subsection with a brief bio sketch and outline of how I practiced reflexivity throughout the study. I am an international PhD student in Greensboro, North Carolina. I was born in Lesotho at the turn of the Apartheid era in neighboring South Africa. My K-12 equivalent schooling occurred in the insular environment of international schools modelled on and affiliated with Cambridge IGCSE and the International Baccalaureate Organization's IB program. The idealism of education in a well-resourced international schooling system while surrounded by an under-resourced public school system while living in a majority Black country meant that my awakening to the harsh reality of being black in a world where Western cultural and English language hegemony was delayed to my undergraduate studies in South Africa. My belief in meritocracy changed at this time as the effects of the disparities created by Apartheid on college students along both race and class lines.

This social consciousness grew as I worked on my Master's dissertation, which focused on the role of language in mathematics teaching in public school classrooms. I observed mathematics teachers switch between languages for different purposes, listened to them explain their reasons for switching and witnessed their struggle to balance the cognitive benefits of teaching in a student's home language and the pressure from administrators to teach in English.

I was drawn to the present research topic because I interacted with course materials in the measurement classes that communicated implicit hierarchies in educational attainment that seemed to be coded. For example, the certainty that differences in achievement scores across urban and suburban schools or Catholics and public schools were articulated by instructors and not questioned by my peers led me to realize that there are coded ways in which race operates, and I was not privy to the key. My position as an international student, socialized in a different part of the world, was advantageous throughout the study because I could question what is taken for granted.

As a Black woman who is a mother to a daughter in the K-12 schooling system, I acknowledge that I am susceptible to sweeping statements that inflate the effect of race or ethnicity on constructs such as academic achievement, behavior patterns, and so on. This was particularly challenging when preparing for and conducting interviews with my research participants. A key part of my interview preparation was doing light background research on the research teams, their professional affiliations, research agendas, labs, and special centers they are part of, and so on. All the researchers interviewed were white, middle-aged males who were well-published in their respective fields. Given the racial tensions prevailing in US society, it is difficult to have conversations about race across races. I was particularly mindful of the potential

desire to be politically correct and non-offensive by the researchers I interviewed. Similarly, I was mindful of not drawing them into conversations beyond their comfort level.

To this end, I kept a reflective journal and created space in my data collection tools to document 'knee-jerk' reactions that I revisited later and engaged colleagues and professors of different racial and ethnic backgrounds in reflections on the research studies, the quantification of data, the conceptualization of race and ethnicity, and preliminary findings. I believe that writing down thoughts, feelings, and beliefs makes it easier to assail and interrogate by myself and in conversation with others, such as faculty advisors and peers. I also included reflections from interactions that were not explicitly related to my research study but greatly impacted my data collection tool.

CHAPTER IV: RESULTS

Introduction

This study aimed to explore trends in the conceptualization of race and ethnicity as grouping variables in published journal articles that use or include differential item functioning (DIF) analyses. Table 77 below restates the research questions and maps them to the respective phases of the research. As shown in Table 7, research question 1 was addressed by Phase 1, the quantitative phase, research question 2 was addressed by both the quantitative and qualitative phases of the research, while research questions 3 and 4 were answered by Phase 2.

Table 7. Research Questions

RQ		Phase 1	Phase 2
1	What are key characteristics of the differential item functioning investigations that employ race and/or ethnicity as grouping variables and appear in peer reviewed journal articles published 2015-2020 literature?	x	
2a)	What trends emerge in DIF analyses that use race and/or ethnicity as grouping variables in terms of **how the terms are defined**?	x	x
2b)	What trends emerge in DIF analyses that use race and/or ethnicity as grouping variables in terms of **how the categorization is conducted**?	x	x
2c)	What trends emerge in DIF analyses that use race and/or ethnicity as grouping variables in terms of **how findings are reported and interpreted**?	x	x
3	How does the conceptualization of race and ethnicity in recently published DIF studies respond to the particularities of the study contexts?		x
4	To what extent do findings from DIF studies that use race and/or ethnicity as grouping variables align with the authors' they cite in terms of interpretation when utilized in future research?		x

The findings presented in this section relate to the dataset obtained by a systematic search of published peer-reviewed journal articles from five academic databases (APA PsychInfo, ERIC, Education Source, Business Source, and CINAHL), which contained key terms 'DIF' or 'differential item functioning' AND 'race*' or 'ethnic*' for the period 2015-2020. Based on this search criteria, 177 unduplicated articles were identified, and after screening of abstracts, 128 articles were included. Upon review of the manuscripts, 8 were further excluded, resulting in 120 articles that met the inclusion criteria.

Research Question 1: Key Characteristics of DIF Investigations

This section presents findings on the key characteristics of DIF investigations that used race and/or ethnicity as grouping variables in response to Research Question 1. These investigations are drawn from 120 journal articles published in peer-reviewed academic journals. I begin with characteristics related to the publication channels, including year of publication, field of study, study location and journals. The section will then report on trends in study details, including the types of tests studied, study sample sizes and details of datasets used and conclude by presenting trends related to the DIF analyses performed, including DIF conceptualization, DIF detection methods employed, and grouping variables used.

Publication Channels

A breakdown of the year of publication, field of study and study location are provided in Table 8 8 below. In terms of year of publication over the six years considered, there was an equal distribution of published articles of DIF analyses that use race and/or ethnicity. In 2016 and 2018, there were particular issues of '*Psychological Test and Assessment Modelling'* and '*Quality of Life Research'* respectively that related to the validation of various Patient Reported

Outcomes Measurement Information System (PROMIS) instruments, which accounted for the high incidence of articles with DIF analyses published in those years.

The field of study was determined as part of the review of each article. Articles were considered to be in the **public health** field if they related to tools/instruments used in clinical settings (e.g., positive aspects of caregiving, kidney disease quality of life, family satisfaction with end-of-life care, self-reported memory problem scale, food parenting practice, etc.) and/or related to public health concerns that were not strictly psychology/psychiatry related (e.g., HIV disclosure belief scale, cannabis use disorders identification test, etc.) Articles were considered to fall into the **psychology/psychiatry** field if the tools/instruments were related to psychological constructs (e.g., religiosity, empathy, anxiety, psychological distress, hardship, depression, etc.). These two categories accounted for most articles in the sample, with 43% being in the public health field and 38% from the psychology/psychiatry field. Only 17% (n = 20) of studies were in the field of education, which lends credence to the acknowledgement that the use of DIF analyses is no longer the preserve of educational testing (Zumbo, 2007).

Regarding study location, most studies reported in articles were conducted in the USA (63%, n=76). Studies in Singapore contributed 7% of the articles, while those in the Netherlands contributed 4%. At the same time, South Africa, Canada, Malaysia, and the UK/England each had a 3% representation in the study sample. These results suggest that DIF analyses that use race and ethnicity as grouping variables are more prevalent in the United States, with smaller but notable contributions from Singapore, the Netherlands, South Africa, and Canada. Further, 10 % of studies were multinational, i.e., conducted in two or more countries.

Table 8. Overview of Articles Sampled

Year of Publication	**n**	**%**	**Study Location**		**%**
2015	14	12%	USA	76	63%
2016	22	18%	Singapore	8	7%
2017	14	12%	Netherlands	5	4%
2018	23	19%	South Africa	4	3%
2019	24	20%	Canada	4	3%
2020	23	19%	Malaysia	3	3%
			UK / England	3	3%
Field of Study	**n**	**%**	China	2	2%
Public Health	51	43%	New Zealand	2	2%
Psychology/Psychiatry	46	38%	Australia	1	1%
Education	20	17%	Germany	1	1%
Criminal Justice	3	2%	Indonesia	1	1%
			Two countries	5	4%
			Three or more countries	5	4%

Studies authored in Singapore were closely related, with seven of the eight articles affiliated with the Institute of Mental Health Research Division, the Program in Health Services & Systems Research, Singapore General Hospital and other hospitals and hospital departments. A cursory look at authorship showed a massive overlap in authors, suggesting that all seven research articles came from a single research group or agenda. In addition, these articles all studied health-related quality-of-life diagnostic measures in adults and seniors and included the DIF analyses to validate said measures. The remaining article was in the education sector, was solo-authored, and was less likely to be part of the research agenda.

Similarly, articles published in the Netherlands were closely related to a core team of three authors on four studies that used the same dataset (See Datasets section below).

Journals

Table 9. Number of Articles Published per Journal

Journal	N	%
Quality of Life Research	11	9%
Psychological Assessment	8	7%
Psychological Test and Assessment Modeling	8	7%
PLoS ONE	4	3%
Frontiers in Psychology	3	3%
Journal of Applied Measurement	3	3%
Appetite	2	2%
Assessment	2	2%
Cultural Diversity & Ethnic Minority Psychology	2	2%
Health & Quality of Life Outcomes	2	2%
Journal of Criminal Justice	2	2%
Journal of Psychoeducational Assessment	2	2%
Malaysian Journal of Learning and Instruction	2	2%
Psychology of Addictive Behaviors	2	2%
Social Psychiatry and Psychiatric Epidemiology: The International Journal for Research in Social and Genetic Epidemiology and Mental Health Services	2	2%
The International Journal of Behavioral Nutrition and Physical Activity	2	2%

Six journals, *Quality of Life Research*, *Psychological Assessment*, *Psychological and Test Assessment Modeling*, *PLoS ONE*, *Frontiers in Psychology*, and the *Journal of Applied Measurement*, were responsible for 31% of published DIF articles with race and/or ethnicity as grouping variables (See Table 99). Two of the journals, *Psychological Test and Assessment Modelling* in 2016 and *Quality of Life Research* in 2018 released special issues that focused on the validation of various Patient Reported Outcomes Measurement Information System

(PROMIS) instruments. Ten journals had two articles each and 63 articles were published in distinct journals, accounting for about 53% of the sample.

Authorship

The distribution of authorship in the dataset follows a pattern with varying degrees of collaboration. Most articles were co-authored, and only three had a single author, a relatively small portion of the dataset (3%). This suggests that solitary authorship, while present, is not the dominant pattern in the sampled articles. A sizable number of articles had between two and eight authors, collectively accounting for most of the dataset (108 out of 120 articles; 90%). This indicates a prevalent trend of collaborative research where authors work together in small to medium-sized teams. Articles with higher numbers of authors (more than eight) are less frequent in the dataset. Of the 12 articles with more than eight authors, five were conducted in Singapore, five in the USA, one in Australia, and one was a multinational study. These instances of extensive collaboration represent large-scale research projects in the medical field and consortium-based research efforts in the case of the multinational study.

In the analyzed dataset, 92 articles had distinct first authors. Eleven authors were identified as the first authors of two articles each. An exceptional case within the dataset was author Teresi J. A., the first author of a remarkable six articles. Table 10 below presents the distribution of authors of the articles sampled.

Table 10. Frequency Distribution of Number of Authors

# of Authors/Article	n	%
1	3	3%
2	9	8%
3	16	13%
4	21	18%
5	19	16%
6	19	16%
7	9	8%
8	12	10%
9	4	3%
10 +	8	7%
Total	120	100%

Study Details

Type of Tests

To comprehensively present research results in this context, it is essential to define the various types of tests subjected to DIF analyses. Table 11 below depicts frequencies and percentages of the different types of tests analyzed in articles in the study sample.

Table 11. Type of Test by Field of Study

	Type of test studied							
Field	**Diagnostic**	**Attitude**	**Proficiency**	**Personality**	**Aptitude**	**Other**	**Total**	**Percentage**
Public Health	38	6	4	-	-	2	**50**	42%
Psychology/Psychiatry	35	-	6	4	-	1	**46**	38%
Education	6	6	6	-	2	1	**20**	17%
Criminal Justice	1	1	-	1	-	-	**3**	3%
Total	**80**	**13**	**16**	**5**	**2**	**4**	**120**	
Percentage	67%	11%	13%	4%	2%	3%		

Diagnostic tests are designed to identify the presence or absence of a particular condition or characteristic, such as a medical diagnosis or educational assessment. Diagnostic tests play a pivotal role in healthcare and education, and understanding potential DIF in these tests is critical for ensuring fair and accurate evaluations. Of the 80 tests subjected to DIF analyses, the articles were classified as diagnostic tests (67%), and 73 were in the public health and psychology/psychiatry field. Diagnostic instruments measured mental health and psychological (e.g., depression, anxiety, positive mental health, feeling tone, etc.); physical health and quality of life (e.g. physical functioning, kidney disease quality of life, fatigue measure, etc.); emotional and behavioral assessment (e.g., pediatric behavior problems, self-efficacy managing daily activities, coping and adaptation processing, etc.); social and relationship assessment (e.g., meaning and purpose, satisfaction with social roles, loneliness); cognitive and memory assessment (e.g., applied cognition, junior metacognitive awareness, etc.); and substance abuse and addiction-related assessment tools (e.g., cannabis use disorders identification, Rutgers alcohol problem index, internet addiction, etc.). Details of the diagnostic tests and instruments investigated are provided in Table 12 below.

Table 12. Types of Diagnostic Tests

Type of Diagnostic Test	**n**	**%**
Mental Health and Psychological Assessment	27	34%
Emotional and Behavioral assessment	15	19%
Physical Health and Quality of Life Assessment	15	19%
Social and relationship assessment	11	14%
Cognitive and Memory Assessment	8	10%
Substance Abuse and Addiction	4	5%
	80	100%

Attitude tests, in turn, are employed to measure an individual's beliefs, opinions, or sentiments towards a particular topic or subject. 11% of the articles reviewed studied the DIF of attitude tests (n = 13). About half of these articles were on public health; for example, (van Zyl et al., 2015) studied the HIV Risk Measure in South Africa, and (Hu et al., 2017) studied the HIV disclosure belief scale in the USA. Examples from the education field include (Bowe, 2019), who administered four attitude scales to multiethnic students as part of the Longitudinal Study of Young People in England and (Adams et al., 2020), who studied students' cognitive engagement, emotional engagement and behavioral engagement in a blended learning model of instruction in Malaysia. One article studied an aptitude test, the Neighborhood Disorder Scale, from the criminal justice field (Ward et al., 2017).

Proficiency tests assess an individual's competence or skill level in a specific domain, such as language proficiency or technical skills. In the sample of articles, 8% were proficiency tests. Most were from the education field and related to language proficiency in the US, Indonesia and England (Devine & Hughes, 2016; Farrington & Lonigan, 2015; Sandilos et al., 2015; Tjipta et al., 2019). In the psychology/psychiatry field, (Cordier et al., 2019) studied a Pragmatics Observational Measure (POM), (Aksu-Dunya et al., 2020) administered a proficiency test of Socio-Emotional Learning to K-3 students in the USA, and (Lindhiem et al., 2019) administered a test of knowledge of effective parenting.

Personality tests measured an individual's personality traits, motivations, and emotional makeup and were mostly observed in the study sample in the psychology/psychiatry field. Five articles (4%) studied personality tests out of the 120 sampled. In criminal justice, the Gramsick Scale for Self-Control was studied by (Ward et al., 2018) and in psychology/psychiatry, (Geldenhuys & Bosch, 2020) studied the BEM sex role inventory, (Du Plessis & De Bruin, 2015) studied personality item pools, and finally (Harpole et al., 2015) studied the fear of negative evaluation scale.

Aptitude tests evaluate an individual's potential to acquire specific skills or knowledge. The study sample consisted of two articles that subjected aptitude tests to DIF analyses where

race and/or ethnicity were used as grouping variables from education (Adams et al., 2020; Hasnain et al., 2017).

The **'Other'** category was used for tests that did not fit into either of the five categories listed above and only occurred once or twice. They included two versions of a measure of identity, the Multi-group Ethnic Identity Measure, which was distinct from personality traits (Loyd, 2019 & Chakawa, 2015). The sample also contained one measure of experience, one of adherence to treatment (Lange, 2015) and one of patient experience (Setodji et al., 2019).

Datasets

One of the characteristics collected from each of the 120 articles in the sample was the source of the data set used. The abstract and methodology sections were used to identify how the data was subjected to DIF analyses and are summarized in

Table 133. This variable was collected to ascertain the degree of control authors had over how the race and ethnicity variables were conceptualized and operationalized, as in instances where authors subjected data to secondary analysis, they had little to no control over conceptualizations and operationalizations of race and ethnicity. Data was mainly collected by the researchers (*n*=92, 77%), most of which were collected solely for the study reported in the article (*n*=56), and the rest were collected as part of a larger study. Details of the larger studies are provided in Table 144, along with the number of DIF studies included in the sample. In 28 articles, authors used an existing data set and subjected it to secondary data analysis.

Table 13. Details of Datasets Used in the Studies

Dataset that was used	n	%
Author collected solely for this study	56	47%
Author collected as part of a larger study	36	30%
Existing dataset	28	23%

Table 14. Details of Larger Studies

Study	Description	Articles
PROMIS - Patient Reported Outcomes Measurement Information System (N=18)	An NIH-funded program that develop and validate person centered self- or parent-reported measures and question banks to evaluate and monitor physical, mental, and social health in adults and children in the general population and those living with chronic conditions.	Jones et al *(2016)*; Fieo et al *(2016)*; Reeve et al *(2016)*; Teresi et al *(2016a)*; Teresi et al *(2016b)*; Jensen et al *(2016)*; Hahn et al *(2016)*; Quach et al *(2016)*; Teresi et al *(2016c)*; Hong et al *(2016)*; Salsman et al *(2019)*; Salsman et al *(2020)*; Terwee et al *(2019)*; Rose et al *(2018)*; Silverberg et al *(2020)*; Tucker et al *(2020)*; Forrest et al *(2018a)*; Forrest et al *(2018b)*
HELIUS - Healthy Life in an Urban Setting (N=4)	A large prospective study carried out by the Academic Medical Centre at the University of Amsterdam and the Municipal Service of Amsterdam. A random sample of nearly 25 000 participants was selected using stratified random sampling in 2011-2015 from the municipal	Miller et al *(2019)*; Galenkamp et al *(2017)*; van Amsterdam *(2019)*; Galenkamp et al *(2018)*

	registry. The purpose of the study was to uncover the reasons behind the disparities in disease prevalence among the major ethnic communities in Amsterdam.	
LSYE - Longitudinal Study of Youth in England (N=2)	A longitudinal study that follows the lives of around 16,000 people in England born in 1989-90.	Bowe *(2017; 2019)*

Study Sample Size

Specific guidance on the required sample size for DIF detection varies depending on the method used. In general, investigations that use the Mantel-Haenszel approach to DIF detection recommended a sample size of 200 - 250 examinees per group (Clauser & Mazor, 1998), while IRT-based DIF detection methods require large samples for accurate parameter estimations for each of the compared groups depending on whether a two-parameter (2PL) or three-parameter (3PL) model is selected.

Sample size varied widely among the articles reviewed (min = 140, max = 44 846) with an average of 3, 855 examinees (SD = 7 053). The distribution of sample sizes in articles is provided in Table 15. There are no hard and fast rules relating to sample size designations in the literature and educational measurement textbooks; thus, the categorizations provided in Table 15 were assigned for ease of reference, not to be instructive or prescriptive. A fifth of the articles included in this systematic review (*n*=24) had sample sizes that were very small (less than 500 participants). Forty percent (40%) of the articles sampled included study sample sizes between 1, 000 and 5, 000 participants. Only eight articles (7%) had substantial sample sizes of more than 10, 000 participants. As shown in **Error! Reference source not found.**6 below, a contingency t able of test length by sample size, these articles with very large samples tended to study short tests (less than 20 items).

Table 15. Summary of Sample Sizes in the Articles

Category label	N participants	N	%
Very Small	< 500	24	20%
Small	500 - 1 000	20	17%
Medium	1 000 - 5 000	48	40%
Large	5 000 - 10 000	17	14%
Very Large	> 10 000	8	7%

Table 16. Contingency Table of Study Sample Size by Test length

Sample Size		Test Length (# items)				
		Very Short (<10)	Short (11 - 20)	Medium (21-50)	Long (51-100)	Very Long (>100)
Very Small	< 500	5	6	4	3	2
Small	501 - 1 000	4	9	9	1	1
Medium	1 001 - 5 000	13	14	23	1	0
Large	5 001 - 10 000	8	5	1	1	2
Very Large	>10 000	2	4	2	-	-

Of the four articles, two distinct populations were sampled within the same study. For instance, in (DuPaul et al., 2020), the same instrument was administered to teachers and parents; in (Ning, 2018), two versions of the same instrument - 18 items and 12 items - were administered to different samples of students and in two studies (Forrest, Devine, et al., 2018; Forrest, Ravens-Sieberer, et al., 2018) administered the same instrument to children and parents.

DIF Analyses

DIF Conceptualization

Details of the study conceptualization for each journal article were collected using the data quantification form described in Chapter 3 in line with the five purposes of DIF analyses proposed by (Zumbo, 2007) with the addition of a sixth category, 'validation of a new instrument' which is a special case of 'investigating threats to internal validity'. The categories are conceptually very related but were coded to be mutually exclusive. Frequencies and percentages of articles in each category are provided in Table 17 below.

Table 17. DIF Study Conceptualization

DIF conceptualization	**n**	**%**
Investigating threats to internal validity	33	28%
Validation of a new instrument	28	23%
Establishing measurement invariance	27	22%
Understanding item response processes	18	15%
Comparability of translated/adapted tests	8	7%
Fairness & equity in testing	6	5%

In 33 studies, DIF was conceptualized as a form of investigating **Threats to internal validity** (28%) with the articles stating in the methodology section or elsewhere that DIF analyses were conducted as a precursor to group comparisons formed using race and/or ethnicity. Some examples of the purpose/goal of the articles are provided below:

> DIF analysis should be performed before comparing composite scores across subgroups of other sampled populations. (Janulis et al., 2018)

> Before making clinical decisions based on evidence of differences in HRQOL between Black and White patients, the measurement invariance of HRQOL measures between Blacks and Whites need to be established. (Peipert et al., 2018)
>
> In order to conclude ethnic differences in the prevalence of depressive symptoms, one should verify whether items of a depression questionnaire measure the same concept in all groups, i.e. confirm that the questionnaire is measurement invariant. (Galenkamp et al., 2017)

Validation of a new instrument was another leading conceptualization for DIF studies reported in the peer-reviewed journal articles included in the sample, with 23% of articles in the studies (*n*=28) stating that the major goal or purpose of the study was to create or otherwise validate a test or test item bank. For instance, (Kwan et al., 2019) "developed a comprehensive and culturally sensitive PM [positive mindset] item bank to measure PM in Singapore." In ten of these studies, the primary or secondary aim was to select, based on DIF findings and through other item measures such as item-subscale correlations, discrimination values, and a good fit to IRT models, optimal items for use in final or shortened measures. For example, (Stone et al., 2020) had the goal of shortening the [Children's – Power of Food Scale] C-PFS to 15 items "with a goal of retaining reliability, strong inter-item relationships, and good coverage of levels of hedonic hunger."

Establishing **measurement invariance** was identified as the conceptualization for DIF studies reported in the peer-reviewed journal articles, 23% of the articles included in the sample (*n*=28) stating that the major goal or purpose of the study was to establish measurement invariance and made mention of this analysis as a means to establish that items function similarly for different demographic groups or a precursor to model-based statistical measurement frameworks such as IRT. In this way, DIF is the empirical, methodological approach to

determining a) lack of invariance, b) model-data fit, and c) model appropriateness (Zumbo, 2007). Below are some examples of this conceptualization of DIF.

> This study aimed to investigate whether the factorial structure of the [Fagerström Test for Nicotine Dependence] FTND is similar across five different ethnic groups monitored by the HELIUS study group. (van Amsterdam et al., 2019)

> This study aims to identify whether features of psychopathy commonly assessed in youth can be measured consistently across childhood and adolescence and equivalently between African American and Caucasian youth. (Hawes et al., 2018)

> The present study applies Rasch analysis to SELweb's [emotion recognition] ER assessment to evaluate SELweb ER's psychometric properties, including dimensionality, item fit, and DIF by gender and ethnicity. (Aksu-Dunya et al., 2020)

> Latent variable modeling, including factor analyses and IRT, was used to examine the psychometric properties of the [Mediation Management Test] MMT and the resulting short-form, including dimensionality, reliability, information, and measurement equivalence. This is the first application of IRT to examine a medication management test. (Teresi et al., 2018)

In 18 articles, the DIF analyses were employed to **understand item response processes**. Articles were put into this category if the study purpose or research specifically mentioned of the construct under investigation and the findings of DIF in and of themselves are hypothesized to be findings. This is demonstrated in the following three article excerpts:

> It is crucial to verify the extent to which mental health scales used in the United Kingdom's national surveys are culturally sensitive, not only to demonstrate the reliability of the scores and validity of inferences made for all ethnic groups but also, **more importantly, because findings from such studies are used to inform mental**

health policy (e.g., see the work of Barnes et al., 2011).” (Bowe, 2017) *(emphasis added)*

> This paper focuses on methodological issues that arise in measuring quality-of-life domains that include qualitative expressions of positive and negative affect. Specifically emphasized are methods to examine measures that include both positively and negatively worded items and methods used to examine the performance of such measures across groups that differ in ethnic and racial composition.” (Teresi et al., 2017)

> We investigated the differences in the patterns of bilingualism for four different ethnic groups in Indonesia. We conducted DIF to test whether adolescents would have different types of words they only know in one language. (Tjipta et al., 2019)

Only eight articles, representing 7%, used DIF analyses to establish the **comparability of adapted or translated tests**. Articles included in this category made specific mention of translated/adapted measures (e.g., the simplified Chinese version of the Toronto Empathy Questionnaire in (Xu et al., 2020) or the Turkish version of the Patient Health Questionnaire PHQ-9 in (Reich et al., 2018)).

Lastly, six articles were categorized as employing DIF analyses to investigate fairness and equity specifically. Articles were included in this category if, for the purpose of the study, the authors sought to establish an unbiased assessment in relation to race and/or ethnicity or presentation of findings. Considering that the field has moved away from the word bias, particularly concerning race and ethnicity as grouping variables, using this word was considered to align with the fairness and equity conceptualization of DIF. The following two excerpts demonstrate:

> The goals of this study were to (1) evaluate the degree to which the ECog provides an unbiased assessment of functional abilities across three ethno racial (ER) groups of older adults [non-Hispanic White (NHW), Hispanic, and Black] and” (Filshtein et al., 2020)

DIF was tested to ensure that the meaning of what was measured was the same across race, language, and sex. Therefore, we wanted to determine whether the items were biased against any group." (Geldenhuys & Bosch, 2020)

DIF Detection Methods

The DIF detection method used in each article was recorded. The following section reports on the trends in DIF detection methods researchers utilize in empirical studies that use race and/or ethnicity as grouping variables in DIF analyses. I begin by describing the number of DIF detection methods used in each article and report on the frequency of each DIF detection method/approach. The findings for the frequency and percentage of the number of DIF detection methods employed in studies sampled are presented in Table 18 below.

Table 18. Frequency and Percentage of Number of DIF Detection Methods Used

# of DIF Methods Employed	Freq.	%
1	96	80%
2	23	19%
>2	1	1%
	120	**100%**

Many studies in the sample (80%) only used one approach to DIF detection. Nineteen percent (19%) of studies used two approaches. In contrast, only one study from the PROMIS program used six approaches to DIF detection, namely IRT log-likelihood ratio tests, ordinal logistic regression, DFIT, SIBTEST, and MIMIC.

IRT-based approaches to DIF detection were by far the most popular approaches to DIF detection in the studies sampled (marked with an Asterix in Table 19 below), with 68 articles

employing IRT-based DIF detection methods. The easy access to IRT software and open-source software such as R has increased the ease of access to IRT DIF detection methods to authors. Of those, the Lord Wald test was used most often, which tests for the difference in maximum likelihood estimators (MLE) of the difficulty parameters for the two groups being compared (Thissen et al., 1993). It was used in 31 of the 120 studies (26%).

The following popular approach to DIF detection was a non-IRT method, the Ordinal Logistic Regression (OLR). OLR models the log odds of the probability of getting an item correct as predicted from sum correct scores, group membership, and an interaction term of the two, and was used in 30 articles. Given that it is a simple-to-use method for DIF detection (Healy, 2006) that can be used with sample sizes as small as 200 in each group (Camilli & Shepard, 2022), this high incidence of use in DIF detection articles is not surprising.

Multiple Indicators Multiple Causes (MIMIC) models apply confirmatory factor analysis (CFA) to item response data to examine the relationship between the latent trait and background variables and have been shown to be as effective at detecting DIF as more traditional Mantel-Haenszel, likelihood tests and SIBTEST (Finch, 2005). MIMIC models were utilized in 23 cases for DIF detection. Log likelihood ratio tests which compare the fit of response data to models where the parameters of the IRT model of choice (1PL, 2PL or 3PL) are calibrated for all examinees with the parameters constrained to be equal with the model calibrated with the parameters allowed to vary for Group 1 and Group 2 for the item under investigation (Bandalos, 2018; de Ayala, 2009; Osterlind & Everson, 2009). This DIF detection method was used in 17 articles in the study sample. These four DIF detection methods, Lord Wald, OLR, MIMIC and Log-likelihood, account for almost 70% of the DIF detection methods used in the articles sampled.

Surprisingly, the Mantel-Haenszel (M-H) test was used in only 6 articles that formed the sample. The M-H method is a non-IRT method based on the chi-square test. It involves comparing the proportions of correct responses for each group and assessing whether these proportions significantly differ. M-H is reported in the literature as widely used by Bandalos (2018) due to the minimal assumptions and moderate sample size requirement. It is also used by Educational Testing Service (ETS), one of the largest and most prominent standardized testing organizations globally; thus this very low use (5%) was unexpected.

Of the articles that employed more than one DIF detection method, Table 20 below outlines the mix of DIF detection methods according to whether both methods were IRT-based, non-IRT-based or a mixture of IRT and non–IRT-based. Authors generally combined IRT and non-IRT methods (n = 16, 67%). Lord Wald's tests were mostly combined with OLR. Five articles used two non-IRT-based DIF detection methods (e.g., OLR and MIMIC or OLR and CFA), while three articles used two IRT methods together (e.g., Lord-Wald tests and ICC comparison).

Table 19. Most Used DIF Detection Method

DIF Detection Methods	Freq.
*Lord Wald Test	31
Ordinal logistic regression	30
MIMIC	23
*Log-Likelihood Ratio Test	17
*IRT unspecified (incl. Rasch models)	9
*Hybrid IRT Logistic Ordinal Regression	6
Mantel-Haenszel	6
CFA	5
*ICC comparison	5
ANOVA	3
SIBTEST	3
DIF Contrasts	2
DIF Plots	2
Bayesian SEM	1
Difference in Logits	1
DIF logistic technique	1
DFIT	1
Item mapping	1
PROC GLIMMIX	1

* denotes IRT based DIF detection methods

Table 20. Type of DIF Detection Methods Used

DIF Detection methods	N	%
IRT and non IRT	16	67%
Both non IRT	5	21%
Both IRT	3	13%
	24	**100%**

Grouping Variables

Recall that all the articles included in the study sample used race and/or ethnicity as grouping variables. Table 20 summarizes the number of grouping variables used for DIF analysis in the articles sampled, while Table 21 provides a count of the other grouping variables used. 24% of the studies used only race and/or ethnicity as a grouping variable, 19% or 23 studies used race/ethnicity and one other grouping variable, and 23% used two additional variables. In comparison, 24%, or 29 articles, used race/ethnicity and three mother grouping variables. Very few studies used more than five or more grouping variables, including race/ethnicity (11%).

As shown in Table 22, gender, in its binary Male/Female form, was the most common additional grouping variable, as it was used in 78 articles. Age was used in 53 articles, and education level was used in 21 articles. In 16 articles, language was used as an additional grouping variable in the DIF analyses. In most of these instances of the concurrent use of race/ethnicity and language, the focus was on the language of survey/instrument administration (*n*=11). To a lesser extent, respondents'/test takers' language backgrounds were used.

Table 21. Number of Grouping Variables Used in DIF Analyses

Grouping Variables	Freq	%
Race/ethnicity only	29	24%
Race/ethnicity + 1 other	23	19%
Race/ethnicity + 2 others	27	23%
Race/ethnicity + 3 others	29	24%
Race/ethnicity + 4 others	7	6%
Race/ethnicity + 5 others	2	2%
Race/ethnicity + 6 others	1	1%
Race/ethnicity + 7 others	2	2%
	120	**100%**

Table 22. Other Grouping Variables

Other Grouping Variables	Freq
Gender	78
Age	53
Education Level	21
Language of administration	11
Health Status	11
Income/SES	8
Study sample	7
Language background	5
Geographic location	3
Other	10

Types of other grouping variables included in the ‘Other’ category were marital status, employment status, or were otherwise related to the purposes of the studies, such as drug use, institution type, and weekly hours spent online.

Summary

In this section, I presented critical characteristics of the DIF investigations sampled. Concerning publication channels, there was an equitable distribution of DIF analysis articles based on race and/or ethnicity over six years, with notable peaks in 2016 and 2018 due to particular issues in Psychological Test and Assessment Modelling and Quality of Life Research. Most articles are in public health (43%) and psychology/psychiatry (38%), while only 17% pertain to education. Geographically, the USA dominates with 63%, followed by Singapore (7%), the Netherlands (4%), and 3% each from South Africa, Canada, and China, indicating a U.S. prevalence in DIF analyses studies with limited representation from other countries. Six prominent journals, including Quality of Life Research and Psychological Assessment, accounted for 31% of DIF articles with race and/or ethnicity as grouping variables. The findings also revealed a predominant pattern of collaborative research, with 90% of articles having two to eight authors, indicating small to medium-sized teams as the norm; solitary authorship is rare (3%), and instances of more than eight authors, often associated with large-scale medical or multinational studies, are less frequent. Furthermore, 92 articles feature distinct first authors, with one author, Teresi, J. A., standing out with six articles.

In relation to study details, the results presented revealed that in 75% of DIF analyses that use race and/or ethnicity as a grouping variable, the instruments were classified as diagnostic; in 77% of the studies, researchers collected the data themselves, leaving 23% of studies that applied

secondary data analysis to existing datasets. Additionally, 40% of the DIF analyses were conducted using sample sizes between 1000 and 5 000.

The findings related to trends in DIF analyses presented in this section revealed that in 51% of the studies, DIF was conceptualized as a threat to internal validity or as a means to validate a new or adapted instrument. That conceptualization of DIF analyses in terms of fairness and equity was limited (only 5%). Regarding DIF detection methods, 80% of articles used only one DIF detection method; thus, triangulation of findings was limited. Further, IRT-based DIF detection methods were more popular than non-IRT-based methods.

In the next section, I present findings in response to research question 2, specifically related to the conceptualization and operationalization of the race and ethnicity grouping variables.

Research Question 2a: Conceptualization of Race and Ethnicity

As Zuberi (2021) aptly states, "The conceptualization of race is fundamental to all subsequent use of racial data." The nuances surrounding race and ethnicity, both as independent constructs and interrelated phenomena, permeate various facets of research, and understanding their role in psychometric assessments is essential for maintaining scientific rigor and promoting inclusivity. In this section, I present findings relating to race and ethnicity when they are used as grouping variables in DIF analyses. I begin with data on which variable is used and how many articles were sampled, provide definitions for the grouping variable used and conclude the section with the presentation of the definitions provided.

Conceptualization of Race and Ethnicity

Race or Ethnicity

While the terms 'race' and 'ethnicity' are often used interchangeably, it is imperative to acknowledge their distinctiveness. Race typically emphasizes presumed differences seen through observable attributes, while ethnicity is the subjective affiliation based on perceived shared culture, history, language, religion and so on (Kivisto & Croll, 2012). However, these constructs intersect in practice, creating a complex tapestry of identities that researchers must navigate when conducting DIF analyses. In general, the articles considered made no theoretical distinction between race and ethnicity, with the terms used interchangeably even within the same article. While in 24 of 120 articles (20%), race was used as the grouping variable, and 48 articles (40%) used ethnicity as the grouping variable, there was no clear adherence to the distinction drawn in this section (See Table 23). For instance, two articles, Heafner & Fitchett (2018) and Janulis et al., (2018) that used race as the grouping variable included the group Hispanic/Latina/a/x, which in the USA is typically created from the ethnicity census question and even the text of the article referred to race and ethnicity interchangeably. Similarly, articles that use ethnicity as the grouping variable within the body of the article use race and race/ethnicity interchangeably with ethnicity.

Table 23. Race or Ethnicity Variable Used

Variable used		
Race	24	20%
Ethnicity	48	40%
Race/ethnicity	48	40%
Total	*120*	100%

Definitions of Race and Ethnicity

Providing definitions for key terms like "race" and "ethnicity" is fundamental to ensuring readers are clear on what researchers mean by the terms. However, as shown in Table 24 below 78% of the articles sampled failed to provide an explicit definition of race or ethnicity (*n*=94), while only 20 (17%) provided an explicit definition. This low prevalence of clarification on the definition of race and or ethnicity utilized by researchers when reporting their research in journal articles is thought-provoking. It is an invitation to critically examine the implied definitions of race and ethnicity within their respective fields and in psychometrics for those articles that did not provide explicit definitions. A detailed list of articles in each category is provided in Appendix E.

Table 24. Presence of Definitions of Race/Ethnicity

Definition of Race/ethnicity	**n**	**%**
None provided	94	78%
Provided	20	17%
Proxy used and defined	6	5%

Of the articles providing explicit definitions, 13 used ethnicity, four used race/ethnicity, while the remaining three used race as the grouping variable.

Table 25 provides a breakdown of the study location of the 20 articles that provided explicit definitions for race and/or ethnicity. As previously stated, race and ethnicity, as socially constructed variables, have differing histories, trajectories, consequences, and implications that are peculiar to the specific geographic locations where they are situated. Seven studies located in

the USA (7% of all studies in the USA) provided definitions for race/ethnicity; five studies were conducted in the Netherlands, three in South Africa, two from the United Kingdom, two multinational studies and one each from Australia, China, Germany, and New Zealand.

Table 25. Study Locations for Articles which Provide Definitions for Race and/or Ethnicity

Study location	N	Articles	%
Netherlands	5	Galenkamp et al *(2017)*; Galenkamp et al *(2018)*; Miller *(2019)*; Terwee et al *(2019)*; van Amsterdam et al *(2019)*	100%
USA	4	Dmitrieva et al *(2015)*; Kim et al *(2016)*; Parkerson et al *(2015)*; Sandilos et al *(2015)*	7%
South Africa	3	Geldenhuys et al *(2020)*; Loyd et al *(2019)*; van Zyl et al *(2015)*	75%
UK/England	2	Bowe *(2017; 2019)*	67%
International	2	Armenta et al *(2019)*; Goetz et al *(2016)*	20%
Australia	1	Rice et al *(2020)*	100%
China	1	Yang et al *(2019)*	50%
Germany	1	Reich *(2018)*	100%
New Zealand	1	Sandham *(2019)*	50%

Alignment with Official Census Categories

A common practice in the articles under review was to align the definition of race and ethnicity with official population statistics gathering bodies such as the Census Bureau. For instance, Dmitrieva et al., (2015), Goetz et al., (2016), and Kim et al., (2016) aligned with definitions provided by the federal Office of Management and Budget and the US Census Bureau

for self-identified race and ethnicity and the respective census in the UK (Bowe, 2019), in New Zealand (Sandham, 2019), and in South Africa (Van Zyl et al., 2015). One reason for aligning with census categories included the desire for consistency among data collected from multiple research studies; for example, Dmitrieva et al. (2015) consolidated response data to the Center for Epidemiologic Studies Depression scale from four longitudinal studies and Census categories were used for consistency across the studies.

A Case of Netherlands – Alignment with a Longitudinal Study

Interestingly, all the studies in the Netherlands described how ethnicity was determined. Four of the articles used data from the Healthy Life in an Urban Setting (HELIUS) longitudinal study, which is owned by the Academic Medical Center (AMC), and researchers provided the description from the HELIUS study (Snijder et al., 2017). The country of birth of participants and their parents were used to determine ethnicity with clear guidelines of how one was considered to belong to each ethnicity. The articles included the following ethnicities: Dutch origin, South-Asian Surinamese origin, African Surinamese origin, Ghanaian origin, Turkish origin, and Moroccan origin.

Race and Ethnicity as Socially Constructed

Only four articles, three based in the USA and one in South Africa, provided explicit statements of the socially constructed nature of race and ethnicity. In each of these cases, the authors allude to the socially constructed nature of race and ethnicity peculiar to the country the article is conducted in and describe what the race/ethnicity variable represents. Authors incorporated notions of shared language (Loyd, 2019), values, beliefs, cultural practices (Sandilos et al., 2015; Parkerson et al., 2015), shared ancestry and origins (Kim et al., 2016). Two of these definitions highlight the situatedness of the racial and ethnic categories, with Kim

(2016) highlighting that the U.S Census Bureau self-identified racial/ethnic categories denote a country's common social understanding and Loyd (2019) delineating how race and ethnicity function in the context of South Africa.

Description of Multiracial

Two articles described how participants who self-selected multiracial were allocated into the multiracial group. Specifically, Bowe (2019) allocates multiracial participants into the group of the ethnic minority portion of their identity as she leaned on the literature on Biracial identity forming patterns. She states, "For example, I combined White and Black Caribbean biracial adolescents with those who were Black Caribbean, as García Coll and Marks (2012) pointed out that biracial adolescents tend to identify more with their ethnic minority heritage; further society tends to view them as such." This contrasts with Geldenhuys & Bosch (2020), who treated multiracial respondents as a stand-alone racial category, known as 'Colored', in line with common practice in South Africa.

Table 26. Definitions of Race, Ethnicity or Race/Ethnicity Provided in Articles

Definition
Census-based definitions
"In keeping with the specificity of ethnic descriptors for national education data since 2002/2003, the LSYPE denotes 17 ethnic groups but also has a variable that combines the white ethnic groups into one race group and maintains the other minority ethnic groups in their separate categories." **Bowe, 2017** *Ethnicity*
"Note. According to the 2001 Census, Asian Other are typically individuals who are born in either the United Kingdom, Sri Lanka, Middle East, or Africa. Black Other typically consider themselves Black British (Gardener & Connolly, 2005)." **Bowe, 2019** *Ethnicity*

"... and ethnicity groups were created based on New Zealand Census categories."

Sandham et al., 2019 *Ethnicity*

"The sample should be representative of the Dutch general population (maximum of 2.5% deviation) with respect to distribution of age (18–40; 40–65; > 65), gender, education (low, middle, high), region (north, east, south, west), and ethnicity (native, first, and second generation western immigrant, first and second generation non-western immigrant), based on data from Statistics Netherlands in 2016 [36]." **Terwee 2019)** *Ethnicity*

"Race categories similar to those used in official census surveys were used to describe the racial composition of the sample" **Zyl et al. 2015**

"For consistency across studies, we used race and ethnicity data collected by each study to classify participants according to the Office of Management and Budget Standards for maintaining, collecting, and presenting data on race and ethnicity (Office of Management and Budget, 1997), which are also used by the United States Census Bureau." **Dmitrieva et al., 2015** *Race/ethnicity*

"We chose race/ethnicity categories according to published divisions adopted by the U.S. Office of Management and Budget." **Goetz et al., 2016** *Race/ethnicity*

Race/ethnicity defined by place of birth

"Ethnicity was defined according to country of birth of the participants as well as that of their parents (Stronks et al., 2009). This country of birth indicator is the Dutch standard indicator for ethnic origin. It has the advantage of being objective and stable over time, and cross-validation studies showed a high correlation between the country of birth indicator and self-identified ethnic group indicator among Turkish, Moroccan, and

Surinamese people in the Netherlands (Stronks et al., 2009). Specifically, a participant was considered of non-Dutch ethnicity if either of the following criteria was fulfilled: (1) born outside the Netherlands and at least one parent born outside the Netherlands (i.e., first generation); or (2) born in the Netherlands, but both parents born outside the Netherlands (i.e., second generation). A limitation of the country of birth indicator for ethnicity is that people who are born in the same country might have a different ethnic background, which in the Dutch context is applicable to the Surinamese population (Stronks et al., 2009). Of the Surinamese immigrants in the Netherlands, approximately 80% 99econdr of African or South-Asian origin and they were classified according to self-reported ethnic origin. Therefore, after data collection, participants of Surinamese ethnic origin were further classified according to self-reported ethnic origin (obtained by questionnaire) into 'African,' 'South-Asian,' or 'other.' For the Dutch group, we invited people who were themselves, as well as both their parents, born in the Netherlands." **Amsterdam et al., 2019** *Ethnicity*

"Ethnicity was defined according to the country of birth of the participants as well as that of their parents [31]. Specifically, a participant was considered of non-Dutch ethnicity if either of the following criteria was fulfilled: (1) born outside the Netherlands and at least one parent born outside the Netherlands (i.e., first generation); or (2) born in the Netherlands, but both parents born outside the Netherlands (i.e., second generation). In addition, as the Surinamese population consists of different ethnic groups which cannot be distinguished from each other on the basis of country of birth, self-reported ethnicity was used to determine Surinamese subgroups (either African or South-Asian origin). In order to be sure that the respondents report their geographic origin, rather than the group they

feel belonging to, the question on self-identification was phrased in objective terms [31]."

Galenkamp et al., 2017 *Ethnicity*

"Cultural beliefs or expectations about health may lead to differences in the interpretation of specific questionnaire items or in different expectations about health. For example, when asked to rate their health in general, respondents may compare themselves with same aged peers [8]." **Galenkamp et al., 2018** *Ethnicity*

"Participants' ethnicity was defined according to the country of birth of the participant as well as that of the parents (Stronks, Kulu-Glasgow, & Agyemang, 2009). More specifically, a person was defined as of non-Dutch ethnic origin if they fulfilled one of two criteria: (a) they were born outside the Netherlands and had at least one parent born outside the Netherlands (first generation), or (b) they were born in the Netherlands but both parents were born outside the Netherlands (second generation). For the Dutch-origin sample, we invited people who were born in the Netherlands and whose parents were born in the Netherlands. After data collection, participants of Surinamese ethnic origin were further classified according to self-reported ethnic origin (obtained by questionnaire) into African Surinamese, South-Asian Surinamese, Javanese Surinamese, or other/unknown Surinamese." **Miller et al., 2019** *Ethnicity*

We selected three subgroups, differing in ethnicity (no migration background at all vs. Turkish migration background), and language version of the PHQ-9 (German vs. Turkish): Germans with no migration background completing the German version of the PHQ-9 (G-G), Turkish immigrants completing the German version of the PHQ-9 (T-G), and Turkish immigrants completing the Turkish version of the PHQ-9 (T-T). Ethnic groups were

defined by the parents' country of birth according to Schenk et al. [48]. **Reisch et al. (2018)** *Ethnicity*

Differential item functioning, using likelihood ratios, was undertaken to assess for potential differences in responding to APSQ items based on level of education (i.e. potential bias that may occur in differences for those with / without a university degree; Teresi & Fleishman, 2007) and region of birth (e.g. Australian born versus non-Australian born). **Rice (2020)** *Ethnicity*

Race/ethnicity as socially constructed

"Race and ethnicity are both social and politically constructed categories, often used to describe and differentiate individuals and groups of people who share commonalities around physical appearance, historical treatment, heritage, beliefs, language, and traditions (Omi & Winant, 1994), and these categories have been used to define status for various groups. In South Africa, race and ethnicity are interrelated in that people may share racial group membership due to government mandated separation, but they may vary in ethnic group membership (Marx, 1998). In this article, we use terms ascribed to race (e.g., Black and White) as they are typically understood in South Africa due to apartheid separation and ethnicity to refer to ethnic and/or cultural groups that may share racial group classification but vary in language, heritage, and traditions (e.g., Sotho, Xhosa, Zulu)."

Loyd et al., 2019 *Ethnicity*

"Defined by the federal Office of Management and Budget and the U.S. Census Bureau, race and ethnicity in the United States are self-identified categories of respondents' origins. Racial/ethnic categories reflect a social definition recognized in country, rather

than defining race/ethnicity biologically, anthropologically, or genetically (Humes, Jones, & Ramirez, 2011)." **Kim et al., 2016** *Race/ethnicity*

"Within the Hispanic population in the United States, the three largest ethnic subgroups are Mexican, Cuban, and Puerto Rican. Individuals within each of these ethnic subgroups have a shared sense of membership as well as shared cultural traditions, beliefs, and values that make them unique from other subgroups (Wolfram, 1991)." **Sandilos et al. 2015** *Ethnicity*

"For the purpose of the current study, self identified ethnicity was used as a proxy for group-shared cultural and social factors (e.g., shared ancestral, social, cultural and national experiences, commonalities in socialization and SES correlates such as education, debts and assets, political power, and marginalization; Kaufman, Cooper, & McGee 1997; Manly & Echmendia, 2007)." **Parkerson et al., 2015** *Race/ethnicity*

Other

"Chinese ethnic minority groups are very diverse, but the total population of each ethnic group is far smaller than that of the Han Chinese. Therefore, it was important to consider the DIF across ethnicity, but only Han vs. minorities in Chinese sample." **Yang et al., (2019)** *Ethnicity*

"Undergraduate students at four universities in Texas, U.S., and two universities in Chihuahua, Mexico, were recruited for the study via flyers, class announcements, or subject pools … self-identified as either European American/White (U.S.), Mexican/Mexican American (U.S.), or Mexican (Mexico). **Armenta et al., 2018** *Race*

> "Colored is the official term used in the South African racial classification system that denotes people of mixed race. We use this term because it is widely understood in South African studies as a means of comparison and is statistically monitored through South African redress legislation. The use of the term does not indicate that the researchers agree with the continuation of this classification." **Geldenhuys and Bosch, 2020** *Race*

Proxies for Race/Ethnicity

In six articles, a proxy for race or ethnicity was used and described/defined *(See Table 26)*. These studies were conducted in two or more countries, in which case the study location (Ehrich et al., 2016; Haroz et al., 2016; Stevanovic et al., 2017; Lange et al., 2016; and Roy et al., 2016) or country of origin (Park et al., 2019) were proxies for ethnicity.

Research Question 2b: Operationalization of Race/Ethnicity

I now turn to the operationalization of the race and/or ethnicity grouping variables in response to research question 2b, which asked about the trends that emerge in how the categorization of research participants into racial and ethnic groups is conducted. Table 27 below provides a comprehensive overview of the methods employed in categorizing the race and ethnicity of participants in the research articles included in the systematic literature review, elucidating the diverse approaches authors take when operationalizing race and ethnicity as grouping variables.

A significant portion of the sample, constituting 30% (n = 36), did not explicitly mention their race or ethnicity allocation method in the article. Ten of these articles leveraged pre-existing data sets, and nine researchers analyzed data collected as part of a larger study; thus, the authors inherited the methodology employed in those larger and/or original studies.

In most articles, constituting 58% of the total sample (n = 70), test takers/research participants self-selected their race or ethnicity. This approach allows individuals to express their identity based on their subjective understanding, emphasizing the importance of self-determination in matters of personal identity; however, even in cases where participants self-select their race and identity, researchers still need to make further decisions to create the final groups used for analysis. A typical example is the consolidation of separate race and ethnicity; in the US context, participants select their race (American Indian or Alaskan Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, White) and in a separate item asked to select their ethnicity (Hispanic or Latino, Not Hispanic or Latino) creating, in theory ten possible combinations but in reality Hispanic ethnicity being combined with either Black or White races to partition Black and White into Hispanic and Non-Hispanic and the four groups being consolidated into three, form Hispanic, Non-Hispanic Black, and Hispanic White (e.g., Lindhiem et al., 2019). Another typical example is the consolidation of ethnicities into racial groups (e.g., Bowe, 2016; Geldenhuys & Bosch, 2020).

In contrast, a smaller proportion of articles, comprising 4% of the total sample (n = 5), had their race or ethnicity assigned by the researchers. Four of the six articles (Stevanovic et al., 2017; Haroz et al., (2016); Park et al., 2019; and Ehrich et al., 2016) were international studies and researchers allocated participants into the proxy ethnicity category based on the country they participated from. Cartwright et al., (20) consolidated secondary data from five non-interlapping studies that administered the tool under investigation (the MacArthur Community Violence Screening Instrument) and allocated participants into the binary Nonwhite and White categories.

Four articles, constituting 3% of the total sample, used the race/ethnicity provided in records collected by a third party, i.e. pre-existing documentation, such as the medical records

(Jones et al., 2016; Peipert et al., 2018; Teresi et al., 2015) and school records (Lambert et al., 2018).

A smaller subset of participants, making up 3% of the total sample (*n* = 4), fell into the "Other" category. In two articles, race/ethnicity was provided by a third party, such as a caregiver in Weisner et al., (2015), a teacher as in Lambert et al., (2018) and parents and teachers in DuPaul et at., (2020). A special application of DIF was served by Flanagan et al., (2020), where the study design manipulated the ethnicity of a 'typical' student, and teachers responded to the same set of items based on the ethnicity of each "student".

Table 27. Allocation into Racial/Ethnic Categories

Race/Ethnicity Allocation	**N**	**%**
Participant Self-selected	70	58%
Researcher Assigned	5	4%
Third party records	4	3%
Other	4	3%
Not mentioned	36	30%
Total	*120*	*100%*

Racial or Ethnic Groups Considered

In this section, I present findings on the racial/ethnic groups created by the grouping variables race, ethnicity and race/ethnicity. Table 28 below provides a summary of the racial/ethnic groups used for DIF analyses in the research articles sampled in alphabetical order.

Asians were represented in 35 articles (29%). In 15 of the 35articles, which were conducted in China (*n* = 2), China and South Korea (*n* = 1), Indonesia (*n* = 1), Malaysia (*n* = 3),

Singapore (n = 8) and the US (n = 1), only Asian ethnicities were considered. In all 15 articles, the grouping variable used was ethnicity. In studies conducted in Indonesia, Malaysia and Singapore, the ethnic groups considered were Chinese, Malay, Indian and others (e.g., Bumputera, Javanese, Saba-Sawak).

The second most popular racial group in the sample was Black/African American, considered in 69 articles (54%), 60 of which were US-based studies. As per the OMB Directive 15, Black/African American refers to "a person having origins in any of the black racial groups of Africa." Four articles from the Netherlands also included participants with African ethnicity (e.g., African Surinamese, Ghanaian, Moroccan, Turkish), which are related to the Black racial group, and such counted in this category. However, in the less clear case of Moroccan or Turkish, the connection is not as straightforward. In the case of three articles from South Africa (Du Plessis & De Bruin, 2015; Geldenhuys & Bosch, 2020; van Zyl et al., 2015), where race was the grouping variable used, Black was used as the reference group. In only one study (Loyd et al., 2019), also from South Africa, all participants were Black, and ethnicity was used as the grouping variable.

Hispanic/Latino ethnic groups were subjected to DIF analyses in 40% of the studies (n = 48) conducted in the USA. Only five of these articles defined race/ethnicity, four of which referred to the OMB, which stipulated as of 1997 in its 'Race and Ethnic Standards for Federal Statistics and Administrative Reporting' that race and ethnicity be collected in separate questions. As will be described in research question 3, there are at least two ways that this combination/consolidation can be conducted: 1) use the ethnicity question to identify Hispanic/Latino respondents and then allocate all remaining participants into their respective racial groups and prefix it with Non-Hispanic as in Non-Hispanic Asian, Non-Hispanic Black,

and so on or 2) use the two variables (ethnicity and race) in turn to perform DIF analyses. The articles did not detail how the two separate items were combined/consolidated to create the Hispanic/Latino group and the other groups. One study (Sandilos et al., 2015) considered ethnic groups that were all Hispanic, namely Cuban, Mexican and Puerto Rican.

Ten articles included a multiracial/multicultural group (8%). In all ten articles, race/ethnicity was self-selected (n = 7), or the allocation method was not mentioned in the manuscript (n = 3). In two cases that were conducted in South Africa (Du Plessis & De Bruin, 2015; Geldenhuys & Bosch, 2020), the official term for individuals of mixed Black and White race is Colored and is used as an official Census racial category, and thus participants self-selected this category. In one study conducted in the UK (Bowe, 2017), participants were presented with multiple ethnic groups and were permitted to select multiple ethnicities. The other 7 articles did not provide details on how multiracialism/multiculturalism was determined, i.e. whether participants were provided with a multiracial category or were allowed to select multiple racial and/or ethnic identities.

Native American participants were involved in 3% of the studies, and Native Hawaiian/Pacific Islander individuals were involved in 6%. The "Other" racial/ethnic category was used in 36 articles (30%) in different ways, including as a catch-all for all participants who do not fall into a particular racial/ethnic group in binary comparisons, for example, in van Zyl et al., (2015) “other” was used for White, Indian and Colored in comparison to Black (See section on Binary Comparisons below). The admittedly heterogeneous “Other” group was used in the DIF analyses. Another way the “Other” category was used in articles was in addition to multiple racial/ethnic groups. In these instances, the “other” category was mostly subjected to DIF analyses (Chen & Zhu, 2017; Chen et al., 2015; Gay et al., 2016), but in some instances, binary

comparisons were run where one racial/ethnic group was compared to all others in turn (Goetz et al., 2016) and in others, it was unclear whether the "Other" group was used in analyses.

The racial/ethnic groups most frequently studied in the articles sampled were White/Caucasian/European, which appeared in *n* =84 articles (70%). This group was the reference group in all but three articles based in South Africa, where the African or Black group served as the reference group.

Table 28. Racial/Ethnic Groups Considered

Racial/Ethnic Group	N	%
Asian	35	29%
Black	69	58%
Hispanic	48	40%
Multiracial	10	8%
Native American	4	3%
Native Hawaiian/Pacific Islander	7	6%
Other	36	30%
White	84	70%

N = # of articles each racial/ethnic group is considered

The articles in the systematic literature review varied in the number of racial/ethnic groups subjected to DIF analyses. Specifically, more than one-quarter of the articles (31 articles, accounting for 26%) only conducted a single binary comparison involving DIF analyses between two racial/ethnic groups (refer to the details below for further information). In 35 studies (29%),

researchers compared response patterns across three racial groups, while in 31 articles (26%), the analysis extended to four racial/ethnic groups. Notably, 20 articles explored DIF across five or more racial/ethnic groups, while two did not provide details of which ethnic groups were compared. In two articles by Tucker et al., (2020) and Devine and Hughes (2016), there was no indication of the racial/ethnic groups used in the DIF analyses. In both articles, no DIF was found from the analyses. (See Table 299 below)

Table 29. Number or Racial / Ethnic Groups Considered

# Racial/ Ethnic Groups	N	%
2	31	26%
3	35	29%
4	31	26%
5	10	8%
6	8	7%
>6	2	2%
No details	2	2%
Total	**120**	**100%**

Binary Comparisons

In cases where only two racial/ethnic groups are considered, the race/ethnicity grouping variable was used to partition the population of respondents in three different ways. In most articles (*n*=14), only one race/ethnicity was specified, and all other respondents fell into a catchall ‘the rest’ group. For instance, Aboriginal vs. Non-Aboriginal, Australian-born vs non-

Australian-born, Blacks vs Others, Han vs Others, Underrepresented minorities vs Non underrepresented minorities, Whites vs. ethnic minorities, and so on. The second way binary comparisons were conducted was that race and ethnicity were used for grouping, but analyses were confined to only two racial/ethnic groups. Many comparisons (*n*=12) were Black vs White, and others were Chinese vs Australian, German vs Turkish, and Chinese vs Korean. Lastly, articles reporting studies conducted in two countries where study location was used as a proxy for ethnicity, only those two ethnicities were compared (Lange et al., 2016 & Roy et al., 2016).

Excluded Races or Ethnicities

Next, I turn to the racial and ethnic groups reported to be excluded. Of the 120 articles, 35 reported excluding participants belonging to some racial/ethnic groups. The main reasons for exclusions were the research question and hypothesis focus that were narrowly focused on certain racial/ethnic groups and the limitations of participants from certain racial/ethnic groups. In most cases, authors cited a small sample size to justify the exclusion of other racial or ethnic groups. While this is a common reason for the exclusion of racial/ethnic groups when it is the same racial/ethnic groups being excluded, and researchers do not design their studies to intentionally be inclusive, this may result in the erasure of said racial/ethnic groups and continued lack of knowledge and insight into items that function differentially in their favor or to their disadvantage. Typical excerpts from the methods sections of articles state:

- "Data from participants who did not identify with one of the four groups (20% of the broader sample) were excluded from the analyses because no other cultural group was large enough to include in the analyses." (Parkerson et al., 2015)
- "Tsonga, Pedi, Ndebele, Swati & Venda (250) were excluded because the sample sizes were not large enough (<100) for analysis" (Loyd et al., 2019)

- "We excluded respondents who did not belong to the six largest ethnic groups (n =586)." (Galenkamp et al., 2018)
- "Other biracial (128), Black Other (90), Chinese (122) excluded because sample size > 200. Asian Other (342) excluded because the group was heterogenous." Bowe (2019)

Some studies in their design only considered certain ethnic groups. For example, Reich et al., (2018) only considered German and Turkish participants, Armenta & Cooper, (2018) only considered European American, Mexican American, and Mexican in the study design, while Flanagan (2020) only considered European, Asian, and Indigenous Canadians.

Summary

In this section, I presented findings related to the conceptualization and operationalization of race and ethnicity as they are used as grouping variables in published DIF studies. With respect to which of the constructs, race or ethnicity, are used as grouping variables, the findings showed a need for more theoretical distinction between the two. Most articles (78%) did not define the grouping variable used. The section provided an overview of the 20 definitions provided. It showed that seven articles used census-based definitions of racial and ethnic categories, six articles defined ethnicity by place of birth, and only three articles defined ethnicity or race/ethnicity as socially constructed. Articles most frequently compared two (n = 31), three (n = 35), or four (n = 32) racial/ethnic groups in the DIF analyses. Binary comparisons were either based on participant location, one race/ethnicity vs the rest or focused only on two racial/ethnic groups.

The section also presented various trends of racial and ethnic representation in DIF studies. Asians were represented in 35 articles (29%), primarily focusing on specific ethnicities in studies conducted in China, South Korea, Indonesia, Malaysia, Singapore, and the US

Black/African American representation was prominent in 69 articles (54%), mostly in US-based studies, with a few from the Netherlands and South Africa. Hispanic/Latino ethnicity was analyzed in 40% of the studies (n = 48) conducted in the USA, often without a clear delineation between race and ethnicity. Ten articles included multiracial/multicultural groups, with varying methods for identifying participants' racial and ethnic backgrounds. Native American and Native Hawaiian/Pacific Islander participants were involved in 3% and 6% of the studies, respectively. The "Other" racial/ethnic category was utilized in 30% of the articles, sometimes as a catch all for non-binary comparisons or alongside multiple racial/ethnic groups. The most frequently studied group was White/Caucasian/European, present in 70% of the articles, typically serving as the reference group except in South African studies where the African or Black group was used as the reference.

I also presented findings related to reasons for the exclusion of some racial/ethnic groups, which were mostly limited sample sizes and had a narrow research problem focus.

Research Question 2c: Reporting and Interpretation of DIF Findings

Research question 2c sought to uncover the trends in the reporting and interpreting of DIF findings in recently published journal articles that use race and/or ethnicity as grouping variables. Table 30 below summarizes the number of articles, N, for each categorization of interpretation of DIF results and the percentage. In 31 articles (25%), there was no detectable DIF. In the remaining 89 articles, the interpretation of the findings was categorized into one of four categories.

Table 30. Interpretation of DIF Findings

DIF Interpretation	N	%
No DIF	31	25%
DIF with no interpretation	15	13%
Some interpretation	15	13%
Full interpretation	50	42%
Contextualized interpretation	9	7%

Studies showed at least one item with detectable DIF, but DIF results were presented in results tables, re-stated in the narrative, and were categorized as "**No interpretation**." In general, in the narration of findings, the authors list the items that displayed some DIF. Authors almost always proceed with comments on the minimal magnitude, effect, or impact of DIF, which aligns with the small proportion of items displaying significant DIF. As shown in 30 above, 15 articles (13%) fell into this category. In most of these articles categorized as providing no interpretation of DIF, the directionality (i.e. which racial/ethnic group has a lower or higher probability of endorsing the item) is not mentioned. In two cases, Pedersen et al. (2017) and Martin et al. (2020), the purpose of the DIF analysis was to select optimal items for each measure (list the measures respectively), and as a result, the items which displayed DIF were removed from the measure, and no interpretation was provided.

Articles were categorized as providing '**Some interpretation**' if DIF findings were presented in tables and figures and described in the narrative of the results section but not interpreted in relation to the race/ethnicity grouping variable. There were 15 articles (13%) in

this category. Interpretations in this category were characterized by a list of items with DIF, with some articles mentioning the groups between which DIF was detected but without specific mention of the direction of DIF, i.e. which groups were advantaged or disadvantaged by the express DIF. Another feature of interpretations in this category was the impact of DIF, with some authors going so far as to correct the DIF found and compare group means. Most articles (*n* =50) that had at least one item with DIF fell into the '**Full interpretation**' category, where DIF findings were presented in tables and the narrative and were linked to the grouping variable and the direction of DIF was provided.

A few articles (*n* = 9, 7%) provided an interpretation of DIF findings presented in tables and narrative, linked back to the race and/or ethnicity grouping variable(s) and situated the findings in the broader sociopolitical context of the respective study. Such articles were categorized as '**Contextualized interpretation**.' To illustrate, Wiesner et al. (2015) conducted an exploratory analysis using MIMIC models of whether item scores in the Diagnostic Interview Schedule for Children Predictive Scales (DPS) show uniform DIF as a function of gender and race/ethnicity and in presenting their findings situated them in the broader academic literature on mental health in predominantly Latino communities and the high correlation between race/ethnicity and SES. The following excerpt illustrates.

> …it has been documented that racial/ethnic minority children in the U.S., especially those of Latino race/ethnicity, have high rates of mental health services underutilization (Alegría et al., 2010; Kataoka et al., 2002; Snowden & Yamada, 2005). Some have suggested that this might be the result of racial/ethnic (aka, cultural) differences in parents' decision thresholds guiding whether treatment is warranted for specific mental health problems (Alegría et al., 2004; Bussing et al., 1998; Chavez et al., 2010; Yeh et al., 2005; see also De Los Reyes & Kazdin, 2005). Our finding that the ADHD item "taking

medication for hyperactivity" (Item 3) 7 was less likely to be endorsed for African American and Latino children relative to White children, even when their overall mean level of ADHD symptomatology on the latent factor was held constant, fits well with other research on this issue (Eiraldi et al., 2006; Rowland et al., 2002).

Summary

One-quarter of the sampled articles reported no detectable DIF concerning the race and/or ethnicity grouping variables. Another quarter of the sample presented DIF findings when race and/or ethnicity were used as grouping variables but did not relate findings to the race or ethnicity grouping variables or failed to indicate the directionality of DIF, that is, which groups were favored, and which were not. Over 40% of sampled articles (50 articles) presented full interpretations, linking DIF findings to the grouping variable and specifying the direction of DIF, and 9 articles (7%) provided contextualized interpretations, embedding DIF findings within the broader sociopolitical context of the study. Notably, the interpretation variable was crucial for assessing the extent of communication regarding DIF findings and for selecting authors for interviews in the subsequent phase of the study.

Research Question 3: Conceptualization of Race and Ethnicity Response to Study Contexts

Phase 2 of this study was designed to explain some of the findings from the systematic literature review in Phase 1. Due to constraints imposed by the Institutional Review Board I was restricted to including only articles reporting studies conducted within the United States of America to ensure compliance with ethical guidelines and to safeguard participant confidentiality. I sought to find out how the conceptualization of race and/or ethnicity in recently published DIF studies responded to the particularities of the study context. After data analysis was complete for Phase 1, articles based in the US that provided full or fully contextualized interpretations of DIF findings (n = 36) were identified to be included in Phase 2. In addition, two articles that did not provide a full or contextualized interpretation of DIF findings but defined race and/or ethnicity within the manuscript were also included. Email invitations were sent to the first or corresponding authors of the 38 articles identified in November 2023, and reminders were sent two weeks later. Two authors responded to the invitations and were interviewed as part of Phase 2. Due to the limited number of participants in Phase 2 several supplemental materials were used to explicate the study conceptualization for the two studies (e.g., other articles related to the tool, other articles published by the authors) and the study context (e.g., reporting guidelines by the NIH and the US Census Bureau). This section presents findings as case studies to best present how specific conceptualizations of race and ethnicity reflected in each article responded to the unique research area.

The case study approach was beneficial because it leveraged multiple data sources related to each sampled DIF article allowing for rich data. Including forward citations in addition to authors elaborating on their study formation, other articles by authors, and so on provided an in-depth understanding of the contexts.

Before presenting the two cases it is essential to provide some preliminaries to establish a framework for discussing the research findings. To uphold ethical standards of research and to safeguard the identities of the participants I use the following conventions: Case 1 will refer to the research context of the DIF study reported in Article 1. The measured construct will be referred to as Construct A and the measurement instrument/rating scale will be referred to as Tool A. The first author of Article 1 will be referred to as Author A. Similarly; Case 2 will refer to the research context of the DIF study reported in Article 2. The measured construct will be referred to as Construct B and the measurement instrument/rating scale will be referred to as Tool B. The two co-authors interviewed in Case 2 will be referred to as Author B and Author C.

Case 1

Case Presentation and Context

The United States health sector is characterized by widespread disparities in health wherein many racial and ethnic minority populations experience poorer health, higher incidence and/or prevalence of disease, poorer outcomes related to said diseases and higher mortality from specific health conditions (*Minority Health and Health Disparities Definitions*, n.d.). Access to healthcare and other public goods such as voting, geriatric care, and high-quality schooling is also distributed disproportionately among the US population. In addition, Blacks or African Americans, Hispanics or Latinos, American Indians or Alaska Natives, Native Hawaiians and other Pacific Islanders as racial and ethnic groups are underrepresented in biomedical research (National Institutes on Health, 2015). The National Institutes of Health (NIH) is the primary Federal agency for conducting and supporting medical research through its 27 subdivisions. It provides guidelines and policies on conducting basic and clinical research, including racial and ethnic categorization.

Article 1 was published in 2015 and reported the results of DIF analyses performed on secondary data collected in a clinical trial. The article was published in a prestigious journal with a 2-year impact factor of 6.2 (Journal website) and, at the time of compilation of forward citations, had been cited 152 times, according to Google Scholar. It is one of over 100 articles published by Author A and forms part of a 30-year research agenda dedicated to drilling down on the observed replicable race differences on the construct of interest as measured by Tool A. Author A explained that their research team

> decided to pursue a DIF analysis on this scale to try to sort out what might be a real difference [in summary scores] versus what might be more of an artefact because a few items might be driving it or interpreted somewhat differently between African Americans and Whites.

The data used were from administering a post-intervention, 6-month follow-up battery of scales, including Tool A. The clinical trial included multiple interventions administered to a randomly assigned intervention group. Equal numbers of African American, Hispanic, and White participants were recruited to form part of the study and were randomly assigned to the intervention or the control group. The clinical trial concerned designing an intervention beneficial across multiple racial/ethnic groups.

Tool A is an 11-item attitude scale that invites respondents to respond to statements depicting possible positive experiences in a role more traditionally associated with negative emotional and psychological effects on a five-point Likert scale ranging from 0 for ‘disagree a lot’ to 4 for ‘agree a lot.’

Conceptualization of Race and Ethnicity

As this article was a secondary data analysis of an existing data set, Author A and his co-authors had little control over how race and ethnicity were established. Notwithstanding those constraints imposed on Author A and their research team by the secondary data, their conception of race aligned with the National Institutes of Health (NIH) and US Census Bureau's tabulation policies, which ask race and ethnicity questions separately and allow for people of Hispanic descent to identify as either Black or White (Office of Management and Budget, 1997). Author A described Hispanic descent to mean "happen to report Hispanic or Latino heritage, because their families are from Central America, or South America or the Caribbean or Spain," which aligns with the definition of Hispanic provided by the Office of Management and Budget (Office of Management and Budget, 1997). Supplemental articles reporting on other aspects of the clinical trial described respondent ethnicity as a "design indicator variable" along with sex and relationship to the care recipient.

When distinguishing between race and ethnicity, Author A mentioned that the research team deferred to the "NIH mindset", which posits race as "more of a biological driven thing whereas ethnicity is more of a cultural, where your ancestors from thing and Hispanics can be either black or white." Race is considered to be observable and fixed, and linked to genetic makeup. For example, Author A speaks of someone being "clearly white with blonde hair" and when describing the racial variability among Hispanics, clarifies that "they may be genetically as White as I am." Regarding the construct being measured, the differences in mean scores and the DIF observed in Article 1; Author A ascribes the cultural aspects of race that as possible causes.

> I think it is more social and cultural than biological, at least with the factors that drive these kinds of differences that we see in these measures. I do not think that Blacks have a particular genetic advantage to managing (construct of interest) better; I think it is part of the culture of their families and their expectations and perhaps the previous experiences of hardship that they have been through, just on average, they seem to manage it better. I also think people's expectations are a strong driver.

The distinction between race and ethnicity was further evidenced by the different accommodations made in the clinical trial for the Hispanic ethnic group. The study methodology included bilingual (English and Spanish) recruitment materials and the employment of bilingual and bicultural staff.

According to Author A, the concept of prior hardship is also associated with socio-economic disadvantage or low socio-economic status. Further, when transferring findings to different geographic contexts, Author A offered insight on how race/ethnicity findings translate to the concept of dominant/non-dominant, indicating that mineralization is not a function of numbers but is a function of belonging a non-dominant culture -- in their words; “subordinate.”

Author A’s research team had a female African American scholar whom Author A characterized as the “race/ethnicity expert”. It provided insight and unique expertise during the conceptualization of the study and the explanation of findings. For example, this co-author troubled the operationalization of racial/ethnic groups because the procedures homogenized heterogenous groups, and she provided an alternate explanation for the results. Author A had this to say of his co-author:

> Here is an African American woman who has dealt with (the study context) for a long time, and she rustles at this idea that African Americans (participants) do not experience

> as much [the construct]. She says they experience it, do not report it, and are more culturally prepared to take on caregiving roles within their own families, partly because they do not have any other options.

Another issue related to race and ethnicity raised by Author A and their research team was that race/ethnicity was highly correlated with the study site. The clinical trial recruited participants from five sites nationwide: Birmingham Alabama; Memphis Tennessee; Miami Florida; Palo Alto California; and Philadelphia Pennsylvania. Author A proffered "most of the Hispanic [participants coming] from Miami and Palo Alto. Moreover, even in Miami, they are more likely Cuban or Caribbean caregivers, whereas in Palo Alto, they are much more likely Mexican or Central American caregivers."

As outlined earlier, the study design inherited from the clinical trial considered only three racial/ethnic groups: Hispanic, non-Hispanic Black and non-Hispanic White. In relation to Native Americans, Author A explained that the interventions designed as part of the clinical trial, which they characterized as "white looking", would have been inappropriate. Author A provided the following justification for the exclusion of the Native American category:

> I do know something about Native Americans and their, you know, cultural traditions and feelings of discrimination. They are intense as well. And I think, you know, your general, you know, white-looking intervention probably is not going to fit for that group. So, I think they probably left them out for good reasons.

Author A also surmised that the Asian American group would most likely be included as a fourth group in further intervention studies due to their increasing demographic relevance.

In summary, in Article 1, race and ethnicity were conceptually distinguished, with race being fixed and observable, mostly cultural but at least partly attributable to biological or genetic

features. In contrast, ethnicity was posited as cultural and associated with the Spanish language in the clinical trial. In addition, the race and ethnicity items were combined to create three distinct groups: Hispanic, Non-Hispanic Black, and Non-Hispanic White. Racial and ethnic groups were also admittedly heterogeneous, specifically the Hispanic group, which consisted of Latinos from different South American countries.

How Conceptualization Responds to Context

The conceptualization of race and ethnicity in Article 1 was inherited from the clinical study, which was aligned with the NIH as a funding agency. Census definitions for race and ethnicity were used and presented as two separate items (i.e., race and ethnicity) and the three groups were created by combining responses to both items. This conceptualization of race and ethnicity were crude approximations that served to homogenize heterogeneous groups, especially in the case of the Hispanic group where the country of origin of Hispanic participants was known and differed, e.g., Cuba, the Caribbean, Mexico, and Central America.

Race was also conceptualized as observable, fixed, and partly biological. This conceptualization allowed the DIF findings to consider race as a causal variable. While Author A's explanations of how racial/ethnic differences result in the observed differences in scores on Tool A related to cultural peculiarities, some of his explanations relate to the historical and current marginalization of African Americans in health care. Despite this, there was resistance to reading these broader sociopolitical contextual factors as the cause of manifest DIF and explanations were sought within the African American community itself. When it related to Construct A in general, Author A was more open to attributing the cause of low scores on Tool A to the disjointed healthcare system.

Case 2

Case Presentation and Context

The psychological construct, Construct B, measured by Tool B, begins in childhood and is characterized by two distinct subconstructs, subconstruct X and Y, which can sometimes co-occur according to the Diagnostic and Statistical Manual for Mental Disorders (DSM-5) (American Psychological Association, 2013). Trained clinicians determine a diagnosis for Construct B based on reported incidence and frequency of stipulated symptoms by parents and/or teachers. This has been the practice in the diagnosis of Construct B, as Author B explained that "[clinicians'] decisions typically are tied to what they hear from parents and teachers. That is a key part of the diagnostic evaluation for [construct B]. We do not have a test for it."

According to Author B, there are longstanding differences in total scores on Tool B and subconstructs X and Y based on demographic characteristics of children such as age, race, ethnicity, and gender reported in the literature. The aim of the study reported in Article 2 was to examine whether the items on Tool B functioned similarly across these demographic characteristics.

Article 2 was published in 2020 and analyzed parent and teacher ratings of the behavior of children aged between 5 – 17 on two subscales collectively making up Tool B. According to the journal website, the article was published in a journal with a 2-year impact factor of 3.6. At the time of compilation of the forward citations, Article 2 had been cited 33 times, according to Google Scholar. The article forms part of an over 35-year research agenda dedicated to Construct B among children, adolescents and adults by Author B and their co-authors. The research team that authored Article 2 had been collaborating for over 30 years and creating, revising, and validating various versions of Tool B for 25 years.

The study reported in Article 2 validated Tool B with a nationally representative sample benchmarked on the US Census region, family income level, race-ethnicity, and age-sex of child proportions. Data for Article 2 were collected through online-based national research firms. Tool B consists of two 9-item subscales aligned to the DSM-5 to trace school-based and home-based behaviors in children. Behavior is rated on a four-point frequency scale ranging from 0 = Never or Rarely, 1 = Sometimes, 2 = Often, and 3 = Very Often.

Conceptualization of Race and Ethnicity

A distinct feature of the conceptualization of race and ethnicity in Article 2 is that it is not the race/ethnicity of the respondents (teachers or parents) used as a grouping variable but the race and ethnicity of the child whose behavior is being rated. For parent ratings, the child race and ethnicity identified by the parent was assumed to be "pretty accurate" by both Author B and Author C. Both researchers were unclear about how teachers rated children's race and ethnicity as the study procedures did not provide any stipulations in this regard. While Author C regarded the extent to which a teacher could accurately report a student's race or ethnicity as a potential limitation, Author B regarded it as integral to the study, stating that "really, what we are after is: are teachers who at least believe that they are that the child their rating is black versus white, is there a difference in a way that item functions based on that perception?"

In Article 2, separate DIF analyses were conducted for race and ethnicity. The researchers explained that the separation of race and ethnicity constructs was done to align with common research "convention" and US Census methodology. Both authors did not provide further rationale for the difference in the treatment of the concepts of race and ethnicity. This approach results in double counting individuals in racial and ethnic groups in the two analyses, which confounds any DIF findings.

The race analysis was conducted with only two groups, African American and White. The authors explained that the requirements of IRT analyses limited the race analyses to only the largest racial groups. For this reason, the children whose ratings were excluded from the analysis amounted to 23% for teacher ratings and 14% for parent ratings.

The ethnicity variable divided responses into those based on Hispanic children and those based on non-Hispanic children and, as a result, used all the responses available. The authors noted the heterogeneity explicitly created in the non-Hispanic group but were restricted by small cell sizes. Author B went on to express that these small sample sizes prevented them from really drilling down into whether items function differentially for more specific subpopulations in line with the differences in scores that are ubiquitous in the literature on Construct B, a notion expressed as a limitation in Article 2. Author B stated that

> What we ultimately would love to do, but we did not have the sample size, is to look at intersectionality. So, I will go back to the literature around the symptom dimensions, where black males have the highest score. What we ultimately would have loved to have done is look at the intersection between assigned sex and race … but once you start carving the sample up that way, we did not have the cell size to do that.

The demographic characteristics of the informants (parents and teachers) were not considered in any of the DIF analyses, and the limitations section of Article 2 and both co-authors expressed that this would have been ideal. Specifically, Author C cited unconscious bias as "potentially problematic, particularly for teachers, rating students of varying races and ethnicities." According to both co-authors, this would have been challenging due to limitations imposed by reduced sample sizes resulting from considering more variables at once. Author C stated:

> It would have been interesting to look at the demographic characteristics of the informants, the teacher, white versus black teacher, male versus female teacher, and that kind of thing. We did not have sample sizes to be able to kind of split the sample up into small parts. For instance, there are very few male teachers. That would have been interesting to do. It occurred to us that we wanted to do it, but we could not do it.

To summarize, race and ethnicity were conceptualized as distinct concepts in Article 2, though the authors provided no more justification for it being common government policy and research practice. The race and ethnicity used as grouping variables were those of the children rated and not the informants. To this end, race and ethnicity were provided by the informants. In the case of parents, this was assumed to be accurate, and in the case of teachers. At the same time, possible inaccuracy of students' racial and ethnic identity could be viewed as a limitation on the one hand; in this study, the observed race or ethnicity was what was hypothesized to influence item ratings and be the source of DIF. Additionally, both co-authors expressed a desire for a more intersectional approach to the DIF analyses (e.g., to consider race and gender) to better understand the observed differences in total scores on Tool B.

How Conceptualization Responds to Context

Informant Ratings

The nature of Construct B is such that ratings of the frequency of child behavior are reported by parents in the home environment and/or teachers in the school environment for clinical diagnosis. The race and ethnicity of the child are also rated by the informant, which means race and ethnicity are not self-identified, as is often the research practice. The co-authors' explanation of observed differences in scores and the subsequent need to parse out how much of those differences result from DIF substantiate this use of observed race of the children being

rated as opposed to the more common self-identified race and ethnicity. In this way, the research study design for Article 2 has the potential for a differential rating of a child's behavior by the teacher built into the design. The interpretation of DIF findings supported this conceptualization of the study. For example, when referring to the findings related to age, Author B explained that.

> We found that if we control for the overall [score on subscale X], there were certain items that parents and teachers were more likely to report or were reporting at a higher level for younger kids than older kids.

This interpretation focuses on the informants' tendency to endorse an item based on the demographic characteristics of a child being rated, which gets biased indirectly.

Inclusion of Rater Demographic Variables

Related to the fact that the raters provide the race and ethnicity of the children being rated is that the demographic variables, especially race and ethnicity, are not incorporated in the DIF analyses. The inclusion of the rater's race and ethnicity would further explain the differential rating of symptoms. While analyses incorporating the rater's demographic variables were not possible due to limited sample sizes, both co-authors agreed that the race of raters also affects on how children of varying races and ethnicities are rated.

Intersectional analyses

The desire for more intersectional analyses by both co-authors was in direct response to the findings in the research literature for Construct B, wherein differences in mean scores are observed not for race or gender in isolation but for Black males, Black females, Hispanic females and so on. While such intersectional analyses were not possible due to sample size constraints, the research team is attuned to the differential rating of Construct B symptoms, and this

intersectional approach to race and ethnicity responds or will potentially respond to the study context.

Summary

I presented two cases of DIF articles and their respective study contexts in this section. For each case, I examined how race and ethnicity were conceptualized and presented thematically how the conceptualizations responded to the sociopolitical and historical contexts. I conclude with a side-by-side comparison of the two cases (See Table xx). Both articles examined psychological constructs and were precipitated by well-established differences in scale scores on Constructs A and B by participants from different racial and ethnic groups. In both interviews, authors reported adhering to NIH, Census categories, and common research practices as their guide to conceptualizations of race and ethnicity.

There were notable differences between the two cases. The first is the sources of data analyzed by DIF. Case 1 was a secondary data analysis, while the research team in Case 2 collected response data. This difference speaks to the level of control over the conceptualization and operationalization of race and ethnicity as grouping variables. In Case 1, the research team had less control, while in Case 2 there, was more control. The two cases also differed in their treatment of the two separate race and ethnicity items differed in the two cases highlighting the room for alternate operationalizations of race and ethnicity even in studies guided by similar bodies. While in Case 1, the race and ethnicity items were combined to partition responses into Hispanic, Non-Hispanic Black and Non-Hispanic White, in Case 2, the items were used in two different analyses, one for race and another for ethnicity.

Furthermore, in Case 1, race was conceptualized as observable and fixed. Conversely, in Case 2, the research design addressed the potential for bias in raters' assessments by actively

incorporating the rater's perceptions of the race and ethnicity of the children being rated. In this way, the study design in Case 2 deliberately accounted for the possibility of subjective judgment influenced by racial factors.

Additionally, the two cases differed in explaining of the racial and ethnic groups not included in the DIF analyses. In Case 1, certain racial and ethnic groups, such as Asian and Native American populations, were excluded from the analysis due to logistical and potentially political considerations related more to the clinical trial than the DIF analysis. That is to say, other racial and ethnic groups were excluded prior to data collection. On the other hand, in Case 2, the authors cited sample size limitations as the reason for not analyzing data collected from certain racial and ethnic groups (e.g., Asians). This limitation stemmed from practical constraints rather than a priori exclusion.

Research Question 4: Alignment of Forward Citations with Findings from DIF Studies

In this final section, I turn to the fourth research question concerned with aligning findings from DIF analyses with forward citations. Specifically, research question 4 asks: To what extent do findings from DIF studies that use race and/or ethnicity as grouping variables align with the authors they cite in terms of interpretation when utilized in future research? As described in Chapter 3, forward citations were collected from Google Scholar and those publicly available online or through the UNCG Library were compiled into a Microsoft Word document shared with each author at least five days before the scheduled interview. In the second half of each interview, authors were invited to reflect on the forward citations and provided the opportunity to comment on how their work is being used. It bears mentioning that the authors interviewed did not read the forward citations before their respective interviews. As a result, it was difficult to steer away from the conversational pace of the first half of the interview to

reading and commenting on the forward citations. To augment the limited reflections, authors were sent the forward citation document and invited to reach out via email with further comments and reflections. Neither of the authors reverted in this regard.

At some point during both interviews, the authors and I brainstormed a potential study. In both cases, this brainstorming was a natural part of the conversation and served partly as rapport-building. These interactions were not part of the qualitative data analysis plan outlined in Chapter 3 but did elucidate the authors' conceptualizations of race.

The section will be organized as follows: Findings from each case will be presented independently, like Research Question 3 above. I begin by describing how forward citations were compiled and describe the forward citations. I then present the author's reflections on forward citations, conclude with a description of the brainstorming interaction and connect each to the author's conceptualizations of race.

Case 1

Article 1 was published in 2015, at the time of compilation of forward citations, had been cited 152 times, according to Google Scholar. After the removal of duplicates (*n*=), articles published in foreign languages (*n* =3), dissertations (*n* =3) and those without full-text availability (*n*=9), there were 54 forward citations presented to Author A. Citations were ordered with the most recent forward citation being presented first.

From open coding, it emerged that the forward citations differed in their purpose of citing the target article and how they used it to build an argument or make a point. More than half of the forward citations (*n* = 31) used the target article to evidence racial/ethnic differences in the construct. This use of forward citations aligned with Author A's expectations considering the raging academic debate over said racial/ethnic differences in scores, specifically, the finding that

African American and Hispanic respondents have replicable higher scores on the construct. Author A states, “So I suspect a lot of these citations are just in that spirit of ‘Oh, here is another paper that found that African Americans have higher [scores on construct tapped by tool].’”

Many forward citations (n = 15) used the target article to substantiate the construct measured by the tool outside of racial/ethnic differences. This use of the target article findings was not necessarily the product of the DIF analysis or the race or ethnicity grouping variables. Four forward citations used the target article to establish the tool’s validity. Only one forward citation referred to the DIF methodology. This pattern of forward citations was expected by Author A as part of the “roaring” debate over the construct that has been going on for over 30 years. Author A also expected that the racial/ethnic differences that motivated the study reported in Article 1 would be the subject of academic discourse and, as such, form a sizable portion of the forward citations.

Forward citations varied in their couching of racial/ethnic differences in scores on the construct. For the most part, forward citations that mention differences in scores in terms of race/ethnicity as group membership, a typical example being “overall, African Americans have reported experiencing [the construct] more positively compared to Caucasian [participants].” In two cases, however, race was couched as an independent variable; for instance, one excerpt reads, “[construct] ratings are impacted by [participant] race …” while another reads, “identifying as African American or Hispanic is associated with increased [level of construct].” Author A relates this characterization of race as a variable that can impact construct scores with their belief that race and ethnicity, while partly biologically/genetically based, are mostly culturally defined. In reacting to the excerpts above, Author A cited the cultural expectations, norms, and even stereotypes as drivers for respondents standing on the construct. In that way, the

race/ethnicity variable acts as shorthand for cultural expectations that could even be transferable through being embedded in African American families.

> If [African Americans' standing on the construct] is better, I think it concerns some cultural expectations. I have heard people who have heard families say this, especially amongst women; the older African American matriarch is valued, partly because this is what she knows. Moreover, this is what she does. This is what she wants to do, not only for her mother or her aunt but for her sister and whoever. You do not necessarily have that kind of icon in older White families. So, I think that is cultural.

In additional response to the notion of what the race variable does, the presence of an African American co-author with extensive experience with the construct in African American communities on the research team provided more insight into the DIF findings and the complexity of the race/ethnicity variable. Author A shared that his co-author linked African Americans' standing on the construct with systemic exclusion from public health resources. He also underscored the methodological difficulties of self-report methods.

Further Conceptualizations

During the interview, Author A suggested and built on a spin-off study to dig deeper into the observed differences in construct A scores by exploring the effect of being exposed to Black culture in interracial marriages. Specifically, Author A hypothesised that a White woman in an interracial marriage with prolonged exposure to familial expectations of the Black community and the concomitant communal expectations might display similar patterns of scores on Tool A. Author A attributed the higher average scores on Construct A to various factors such as "the culture of their families and their expectations, and, perhaps the previous experiences of hardship that they have been through" which, in theory, are measurable but suggested a more complex

proxy (White woman in interracial marriage) going on to say “That is a good way to sort out how biological (race) is versus how cultural and sort of learned it is.”

Case 2

Article 2 was published in 2020 and, at the time of compilation of forward citations, had been cited 33 times, according to Google Scholar. After the removal of duplicates (*n* =1), books and book chapters (*n* =3), articles published in foreign languages (*n* =5), thesis or dissertations (*n* =6) and those without full-text availability (*n*=1), there were 17 forward citations presented to Author A. Citations were ordered with the most recent forward citation being presented first.

Forward citations referenced Article 2 in several different ways. Mostly, the findings related to the DIF analyses were cited. In four forward citations, DIF findings related to the race grouping variable, while for six, they related to other grouping variables (e.g., age, gender, etc.). Three articles cited Article 2 as justification for conducting DIF analyses on other measures of Construct B and other psychological constructs, part of a call for culturally appropriate assessments for Construct B. Three forward citations were unrelated to DIF findings; one cited Article 2 to explain its exclusion from a systematic literature review, and another two echoed its design limitations and implications for the diagnosis of Construct B.

As stated in the introduction to this section, Authors B and C did not engage with the forward citations before the interview. As such, their reflections were limited. Both authors noted the varied use of their article and were pleased that most of their forward citations were related to Construct B and other proximate psychological constructs.

Further Conceptualizations

While discussing Author B and C’s research teams’ desire to incorporate the demographic variables of respondents into the analysis, we ended up brainstorming how this

could be done while considering the sample size. The authors were limited by the size of their sample, which prevented them from comparing combinations of the variables (e.g., the ratings of White teachers on Black children compared to White teachers on White children). Specifically, over 80% of teachers in their sample that provided ratings were White; thus, the research team was not "able to carve up black teachers rating black kids, black teachers rating white kids, those two cells would be too small. " Alternate conceptualizations discussed were to code for identical race between rater and child and compare it to dissimilar race. This conceptualization would have been another way to understand more deeply what is causing differential endorsement of the items on Tool B depending on the demographic variables of students.

CHAPTER V: DISCUSSION, CONCLUSION AND RECOMMENDATIONS

This study explored the use of race and ethnicity as grouping variables in research journal articles published between 2015 and 2020. The study intended to answer the research questions:

RQ1: What are key characteristics of the differential item functioning investigations that employ race and/or ethnicity as grouping variables and appear in peer-reviewed journal articles published 2015-2020 literature?

RQ2: What trends emerge in differential item functioning (DIF) analyses reported in recently published research studies that use race and/or ethnicity as grouping variables in terms of

a) how the terms are defined,

b) how the categorization is conducted, and

c) how findings are reported and interpreted?

RQ3: How does the conceptualization of race and ethnicity in recently published DIF studies by the researchers identified for RQ1 respond to the particularities of the study contexts?

RQ4: To what extent do findings from DIF studies that use race and/or ethnicity as grouping variables align with the authors' they cite in terms of interpretation when utilized in future research?

The explanatory sequential mixed methods study consisted of a quantitative systematic literature review of published articles DIF that use race and/or ethnicity as grouping variable. Findings from this phase were used to select cases for a second, qualitative phase of the study wherein two US based researchers were invited for an interview to reflect on their study and some forward citations of their work.

This chapter provides a summary and synthesis of major findings. Theoretical and practical implications are woven into the discussion for each research question. Next, methodological considerations and limitations of the study are discussed. Following this, directions for future research are discussed. The chapter concludes by presenting a positionality statement.

Discussion

So far, the findings have been reported by research questions and kept separate. In this section, the information gleaned from the different phases and research questions is integrated as appropriate to provide a richer depiction of issues related to race and ethnicity as grouping variables in DIF research and provide a fuller portrayal of the DIF research landscape as it relates to race and ethnicity grouping variables. The discussion is organized into four broad themes: definitions of race and ethnicity, race or racism, and voice of color and concludes with a discussion on the methodology.

Definitions of Race and Ethnicity

This research study found that of the 120 sampled peer-reviewed journal articles that used race and/or ethnicity as grouping variables, only 22% provided an explicit definition or explanation of what its authors mean by race, ethnicity, or race/ethnicity (depending on the variable used) including cases where a proxy such as study location was used. This finding was similar to what (Poe, 2009) found when examining racial-ethnic group differences in educational research. The two case studies in Phase 2 were among the 78% that neither defined the grouping variable(s) used nor described how allocation into the various racial/ethnic groups was performed. These authors' conceptions and operationalization of race and ethnicity were described and understood from their interviews, highlighting that research manuscripts published

in peer-reviewed journals do not adequately describe these key details. In Case 1, a secondary analysis of data collected in a pre-existing clinical trial study, readers are referred to other journal articles that more fully outline the methodology, including the operationalization of race and ethnicity. Challenges specific to the race/ethnicity grouping variable, such as the high correlation between study location and race/ethnicity and the country of origin for Hispanic participants, were not described in these other articles but surfaced in the qualitative phase of the current study. Similarly, in Case 2, details of the racial categories presented to the respondents were not listed in the manuscript and, at the time of the interview, were not easily recalled by the authors.

This low prevalence of clear descriptions of race, ethnicity, or race/ethnicity and little to no supporting theorization on how race and/or ethnicity intervene on the constructs being measured makes it difficult to fully appreciate and appraise the methodology. Further, this paucity in articulating the operationalization of race, ethnicity, or race/ethnicity as grouping variables also challenges the replicability of DIF studies. Replicability is a major methodological concern, and findings from Phase 1 highlight widespread lack of …

> "Replicability means that the finding can be obtained with other random samples drawn from a multidimensional space that captures the most important facets of the research design. In psychology, the facets typically include the following: (a) individuals (or dyads or groups); (b) situations (natural or experimental); (c) operationalizations (experimental manipulations, methods, and measures); and (d) time points. Which dimensions are relevant depends on the relevant theory: What constructs are involved, how are they operationalized within the theory underlying the research, and what design is best suited to test for the hypothesized effects? Replication is obtained if differences between the finding in the original Study A and analogous findings in replication Study B are

> insubstantial and due to unsystematic error, particularly sampling error, but not to systematic error, particularly differences in the facets of the design." (Asendorpf et al., 2013)

In Case 2, which involved data collection through an online survey distribution platform facilitated by third parties, as stated earlier, the methods section provided no details of how race and ethnicity were presented to raters. Recall that in Article 2, ratings on Tool B, race, and ethnicity of children were provided by raters: parents and teachers. Thus, the methods section was not sufficient to facilitate replication. For instance, if I were to replicate the study, I possibly would present different racial categories, different rules as to how multiracial children are to be categorized and so on.

Kivisto and Croll (2012) define *race* as "the classification of people based on what are presumed to be differences typically evident as differences in physical differences due to such features as skin color." On the other hand, ethnicity is widely understood to refer

> to clusters of people who have common cultural traits that they distinguish from those of other people. People who share a common language, geographic locale or place of origin, religion, sense of history, traditions, values, beliefs, food habits, and so forth are perceived and view themselves as constituting an ethnic group. (Smedley & Smedley, 2005, p. 17)

The 26 explicit definitions for race, ethnicity, and race/ethnicity provided were described in Chapter IV to fall into five categories, namely those definitions that aligned ethnicity or race/ethnicity with census categories, those that used place of birth as the definition or to operationalize ethnicity, those definition that made mention of the socially constructed and situated nature of ethnicity or race/ethnicity, study location as a proxy for ethnicity and other

definitions. In this section, I discuss census-based definitions and socially constructed notions of race and ethnicity.

Population statistics collected in census surveys have always measured race (James, 2008; Khalfani et al., 2008). Before the 1960s, census enumerators observed and recorded race, suggesting that race was "both self-evident and fixed" (James, 2008). Census categories represent a pre-defined and standard way of allocating research participants/test takers into racial and ethnic categories, as in the example presented in Chapter IV, where one research article used Census categories for consistency across several studies. In Phase 2, both cases mentioned using census-based categories and guidelines provided by the NIH as requirements for NIH-funded research studies. For instance, Article 1 was funded by such a research grant as was the primary source of the data subjected to secondary analysis. This is an example of how authors inherit conceptualizations of race and ethnicity. While the history of Census categories is beyond the scope of this project, the Census Bureau shapes the racial and political climate by defining which categories count and redefining the bounds in line with the changing population composition and political agendas (James, 2008; Khalfani et al., 2008). Thus, Author A and their research team were constrained by funding requirements in their conceptualization of race and ethnicity.

On a related note, the two cases in Phase 2 demonstrated how the ethnicity Census item can be used. In Case 1, the Hispanic group was essentially racialized by the creation of Hispanic, non-Hispanic Black and non-Hispanic White groups. This is interesting, given that there was resistance to this very operationalization in the 1930 US census (James, 2008). In contrast Article 2, used the variables to calculate partition scores. There is ambiguity in how Hispanic ethnicity fits in with racial understandings (James, 2008); thus, this finding of the different ways researchers operationalized the consideration of race and ethnicity even while adhering to the

same guidelines on collecting race and ethnicity data. Furthermore, the distinction in the accommodations made for Hispanic participants in the clinical trial in Article 1, such as Spanish translated materials and culturally trained caregivers, while no such accommodations were made for the non-Hispanic Black group, highlights that the two constructs, race and ethnicity, are different in practice.

Roth's (2016) race dimensions typology provides a lens to differentiate the two case studies in Phase 2. Racial self-classification refers to the respondent's subjective self-identification bounded by pre-defined categories. This was the dimension of race used in Case 1, whereas in Case 2, it was the race that raters (parents and teachers) believed the child rated to be. The typology of dimensions of race can facilitate deeper conceptualizations and, consequently, operationalizations of race and ethnicity by inviting researchers to look beyond race and ethnicity as fixed or natural and to think deeper about which dimension is of interest to their research questions and study context.

Race or Racism

One purpose of this study was to learn and understand what race and ethnicity are as variables in DIF analysis, what they are hypothesized to do and how they are engaged to do so. Literature on the history of race and ethnicity as concepts and their development in tandem with, and at times in service to, racial statistics and a eugenics agenda mean that this task was difficult from the outset. Further, the influence of contextual factors such as the political climate and agendas of governments continually change the definition, boundary and meaning of race and ethnicity. In psychometrics, race and ethnicity are generally under-theorized (Russell, 2024). As described in Chapter II, the Standards for Educational and Psychological Testing make few references to race and ethnicity and treat them in very vague terms; for instance, no definition is

provided nor explanations of how the concepts interact and interface with educational and psychological testing. In addition, the following quote from a prominent DIF scholar, editor of seminal DIF textbooks and author of over 170 published journal articles demonstrates.

> I take racial categories, however determined, as given. This is also the plight of the analyst who runs his or her regressions. For the most part, someone else determines the definition of the race variable, and the analyst has to use the available data. I do not apologize for this superficiality because it is the common superficiality of those who employ race as a variable in their analysis. (Holland, 2008)

Notwithstanding this hands-off stance on the theory behind race and ethnicity, they, as variables, are widely accepted not to be amenable to function as causal variables as they are not usually manipulated in experimental study designs (Holland, 2008; Zuberi, 2001). The forward citations found in Phase 2 revealed that the race and ethnicity variables were sometimes interpreted as causal. This is not surprising since James (2008) cautioned that when race is used without any contextualization (historical, political, economic, cultural, etc.) or explanation, the causal mechanism for observed differences lies in the racial categories themselves. She goes on to caution:

> When race is presumed to *cause* differences in family behavior, test-taking, and psychological well-being- that is, without comment or argument about how or why the experience of race in U.S society may result in different outcomes for individuals who face different racialized experiences – conceptual understanding of race as a fixed characteristic is being promoted. (James 2008)

This assertion by James (2008) implies that the causal mechanism that is responsible for the

observed difference is, in fact, racism is echoed by other scholars with (Gillborn et al., 2018; Zuberi, 2001) arguing that since race is not a "thing", participants/respondents have that can *cause* differences critical race theorists look to racism as the mechanism through which the race variable affects or is affected by the construct.

I would argue that given the low prevalence of definitions of race or ethnicity, the even lower prevalence of definitions that acknowledge the socially constructed nature of race and ethnicity, and the limited number of articles that provided contextualized interpretation of DIF findings that in DIF studies it was racism and not race being operationalized I use the two case studies to show how this can occur in different ways. In Case 1, Author A explained that some of the underlying causes of systematically higher scores on Construct A by Black or African American participants were because that community is currently or has been historically excluded from mainstream health care benefits, leading to a necessity to "band together" and create and become comfortable with alternate arrangements. This reasoning was further invoked when reviewing a forward citation from an article based in a different country with a markedly different racial/ethnic composition. In reflecting on this case and using Article 1, Author A explained that being a minoritized population within a country where the dominant group's culture is privileged is analogous to the US situation. In my reading, the domination of one group by another, evident in a disparate distribution of material resources, is racism. Author A, however, resisted my suggestion that race then acts as a proxy for racism.

In Case 2, I argue that in the part of the study that focuses on teacher ratings of students, the variable can be considered racism (or lack thereof) as the observed dimension of race is operationalized to see if ratings differ systematically for African Americans, Hispanic or White children. The mechanism through which the race variable affects scores on Tool B, is therefore,

the perceived race (by teachers) of children. Another study highlighted in Phase 1 by Flanagan et al., (2020) demonstrates how a similar mechanism for the grouping variable can function. In that study, the research design asked teachers to rate their expectations of a "typical student" with the ethnicity being manipulated (Asian Canadian, European Canadian, and Indigenous). Again, in this case, teachers' expectations differed based on their expectations of students of different ethnicities.

Related to this theoretical linking of the race and ethnicity variable to the application of DIF analysis is the concept of situating findings within the broader context. Poe (2009) states

> When researchers use group categories, they need to situate their analyses of the data within the historical and national contextual frameworks in which these categories have become meaningful and acknowledge the institutional frameworks by which these have become the "official" categories for race-ethnicity. (Poe, 2009)

The findings in Chapter IV show that only nine articles situated DIF findings in the broader research and sociopolitical context.

Voice of Color

An essential tenet of CRT/CRQI and QuantCrit highlighted in Chapter II is the importance of acknowledging and centering the cultural intuition of voices of color, which is constituted by the personal and familial lived experiences (Covarrubias & Vélez, 2013; Gillborn et al., 2018). In particular,

> QuantCrit assigns particular importance to the experiential knowledge of people of color and other 'outsider' groups (including those marginalized by assumptions around class, gender, sexuality, and dis/ability) and seeks to foreground their insights, knowledge and understandings to inform research, analyses, and critique. (Gillborn et al., 2018)

The findings from Case 1 showed the impact and influence of this cultural intuition. Author A described how the insights of his co-author, who was an African American woman, troubled the research teams' conceptualizations of race and ethnicity, provided insight into the mechanisms through which race acts to produce different scores on Tool A and even questioned the ability of measurement tools to tap the constructs for African American respondents.

It is widely accepted that the reasons for DIF are difficult, if not impossible, to discern (e.g., Angoff, 1993; Clauser & Mazor, 1998), but the findings from this study suggest that a multiplicity of perspectives, including in the conceptualization of race and ethnicity might improve the prospects of explaining what is causing DIF. Fundamentally, the inclusion of voices of color as CRT/CRQI and QuantCrit recommended along with cultural intuition and lived experiences, would likely significantly enhance the processes for identifying sources of DIF by illuminating 'blind spots' in and even expanding the constructs being measured. Thus, the reasons for DIF might be easier to assail.

Methodology

I conclude this discussion section by considering the methodology and research methods. An explanatory sequential mixed methods approach was used for this study to provide a comprehensive coverage of current studies of DIF in the first quantitative phase and a focused exploration of rationalization, best practices, and reflections on the use of research findings by a sample of authors of published DIF studies that use race and/or ethnicity as a grouping variable. The intention of integration in a sequential explanatory design is to connect qualitative data and results in Phase 2 to explain the quantitative results in Phase 1 (Creswell & Plano Clark, 2018). Specifically, this study used the quantitative results from Phase 1 to identify the results

(variables) that need to be explained qualitatively. Integration of quantitative and qualitative results also happened in the discussion of results above.

Systematic literature reviews of DIF have been performed in psychometrics (e.g., (Berrío et al., 2020; Gómez Benito et al., 2005)) but none have examined the conceptualizations of grouping variables. This phase had several advantages and challenges. One significant advantage was the ability to look across diverse fields, providing a comprehensive overview of where DIF is utilized. This broad exploration allowed for insights into the prevalence of DIF across different academic domains. Additionally, examining popular detection methods sheds light on the methodological approaches used in DIF studies. Moreover, the investigation revealed the extent of exploratory DIF, particularly in understanding whether race/ethnicity was the sole grouping variable used in these studies.

However, there were challenges faced during Phase 1. Homogenizing categories across different countries proved to be a complex task, especially when categorizing individuals as “Asian”, “Black”, or "White". For instance, terms such as Dutch, European Canadian, white, Caucasian, and Australian were all considered under the umbrella term "White," which provided consistency in the analysis but caused an oversimplification of categories. Applying the critical lens of CRT, CRQI and QuantCrit to work that did not explicitly claim to be critical presented another challenge.

In Phase 2, the methodology presented unique advantages and disadvantages. A notable advantage was the novelty of exploring the perspectives of researchers themselves as the data source in the field of measurement. This approach provided valuable context, allowing a deeper understanding of the authors' views and conceptions of DIF and its use in their respective fields.

Additionally, insights into methodological limitations and reservations, often omitted from manuscripts due to limited space, were obtained.

No studies were found that involved inviting researchers conducting DIF analyses in their studies to reflect on their conceptualization and operationalization of race and ethnicity as grouping variables. And in this regard this study is a methodological contribution. Quantitative research is notorious for foregrounding objectivity and implying that researchers advance universally understood and coherent research standards, this study is novel in that it lets readers into the conceptions that inform and shape one aspect of the DIF analysis, the race and ethnicity grouping variables. Indeed, the findings presented in Chapter 4 demonstrate that there was much variation in how even identical guidance from the US Census Bureau is operationalized. Buttressing findings from systematic literature reviews which are often quantitative with the insights from researchers was unique in the field of educational measurement and can be employed in other studies looking into

Despite these advantages, Phase 2 faced challenges, including limited responses. The scarcity of responses may have restricted the diversity of perspectives, impacting the overall richness of the qualitative data. In addition, potentially rich insights would have been gained if researchers had engaged more deeply with the forward citations we foregone because there was not enough time in the interview to allow them to read and be able to comment on the forward citations.

The mixed methods approach, combining quantitative and qualitative elements, offered several advantages. Identifying populations and samples for future targeted research studies, such as researchers studying DIF in specific tools measuring depression, was facilitated.

Interviews explained quantitative findings, offering insights into why certain groups, like American Indians, might be excluded and the bodies researchers look to for guidelines.

Another notable advantage was the inclusion of forward citations as part of the interview protocol. This was a novel contribution in DIF research as it framed the reflection of researchers on their work and their specific research fields and its peculiarities. Bringing researchers into conversation with the use of their DIF related research findings was yet another way to elucidate their conceptualization of race and ethnicity and how the variables act and intervene in their research field. consequences of their presentation of DIF findings which are influenced by their conceptualization and operationalization of race and ethnicity grouping variables. In addition, researchers' interaction with their forward citations is a new and promising aspect of tapping into the consequences of research. While in this study, conversation forward citations focused on the conceptualization and operationalization of race and ethnicity, broader implications for DIF findings could also be the focus of studies that use forward citations as an interview prompt.

Lastly, the brainstorming process with researchers as part of their reflective interviews allowed conceptualizations of race and ethnicity to be extended and further tested was interesting, suggesting a potential for formalizing this aspect in future studies.

Implications

The findings of this study and the discussion have shown that published DIF articles that use race and ethnicity as grouping variables mostly do not provide definitions of race and/or ethnicity. Further, there is variation in the operationalization of race and ethnicity as they are used as grouping variables, including but not limited to how the two are used. Lastly, researchers do not have strong theoretical links between race and ethnicity and the constructs being

measured by tools. This section provides some implications of the findings and discussion presented above for researchers, reviewers, editors, and measurement graduate programs.

Researchers

I present three broad implications for researchers who use race and ethnicity as grouping variables. The first implication relates to articulating some minimum reporting requirements, including definitions for the grouping variable uses (race, ethnicity, or race/ethnicity) and a theoretic link between the chosen conceptualization and the construct being measured. This can also be achieved by providing a more detailed description of how the study's historical, economic, political, or cultural context interacts with, shapes or is shaped by race and ethnicity. In this way, readers, reviewers, and editors can follow the logic of what underlying and intervening constructs/variables race and ethnicity approximate. In addition, more description of how race and ethnicity data are collected (for example, including details of what categories were presented to participants for self-selected race/ethnicity) and manipulated (for example, what happened when participants selected more than one category? How were race and ethnicity on separate items combined to create groups?) and analyzed (for example which groups were compared).

Related to this notion of minimum reporting requirements, researchers might consider locating themselves and their positionality, their lived experiences of race and ethnicity and identifying assumptions they may bring to the DIF analyses. In addition, a researcher might, based on the findings of this study, engage in planned individual and group reflexive exercises to surface and address any biases or knee-jerk reactions throughout their project.

Lastly, the impact and influence of an African American researcher on the research team in Case 1 might suggest that researchers conducting DIF analyses that use race and ethnicity as

grouping variables to deliberately seek out co-authors with different racial and ethnic lived experiences to enrich the conceptualizations of race and ethnicity grouping variables.

Reviewers and Editors

In addition to the implications listed above, journal article reviewers and editors might be more critical of the conceptualization of race and ethnicity used in DIF articles submitted and reviewed and provided guidance on how. To this end, they might increase their knowledge of different conceptualizations of race and ethnicity, the different dimensions of race, and the appropriate research contexts for each. Reviewers and editors might also encourage and even require researchers to situate their research studies more fully in the sociopolitical and historical contexts, for instance, when reporting findings.

Graduate Programs

The background of the study highlights the lack of direct guidance on how race and ethnicity should be conceptualized and operationalized in DIF analyses. Further, the findings of this study suggest that novice measurement specialists/psychometricians need direct instruction on theorizing race and ethnicity as grouping variables and not presenting them as natural, fixed, and universal. Questions to consider in this regard are: What are these variables expected to do or explain? What are the mechanisms through which race and ethnicity are hypothesized to act on the construct? Are there other ways to assail the constructs for which race and ethnicity act as proxies for? In addition, graduate programs might expose students to the different ways in which race and ethnicity are conceptualized and operationalized. Graduate programs might even demonstrate how different conceptualizations and operationalizations might affect findings. In addition, the different ways DIF findings are reported, as demonstrated in the sampled articles

and forward citations, might suggest that direct instruction is needed on how to report findings when race and ethnicity are used as grouping variables.

Areas for Future Research

Given the results, discussion and implications presented above, I have identified the following areas for future research:

1. *More qualitative data collection for Phase 2.* Engaging more researchers who publish DIF analyses in peer-reviewed journals might provide even more insight into conceptualizing the race and ethnicity grouping variables. Interviewing more people from research teams might also show how negotiations around the conceptualization and operationalization of race and ethnicity within the studies go, the group dynamics involved and so on. Alternately, focus group interviews with researchers in the same or similar fields might surface field-specific opportunities and constraints that will further add to what is known about race and ethnicity in DIF studies.
2. *Re-analysis of data in articles* with suggested 'stronger' conceptualizations of race and ethnicity to test whether the CRT approach makes a difference. For instance, the research study brainstormed by the researchers and me in Case 2, where the race and ethnicity of the raters and children rated, would demonstrate how deeper theorizing of the race and ethnicity variables can provide new knowledge.

3. *A deeper qualitative analysis of the DIF articles,* perhaps using discourse analysis to show how DIF analyses (the conceptualization and operationalization of race and ethnicity and the interpretation of results) build or challenge discourses of race. This would further inform how race and ethnicity-related DIF findings are currently and should be interpreted so as not to sustain deficit-based models of minoritized or otherwise dominated populations.
4. *A broader systematic review* that examined all DIF studies would help determine how prevalent race and ethnicity are as grouping variables in DIF studies, trends in other grouping variables used and the degree of theoretical conceptualization and operationalization of those grouping variables.

Reflexivity

I begin this section with the following quote from Clauser &Mazor (1998): "DIF analyses do not lend themselves to a cookbook approach. Most of the steps require judgement, and most require consideration of other aspects of the test development process" as a departure point for highlighting the subjective role that is played by a researcher in DIF research. The numerous aspects of DIF detection that require judgement, such as the selection of grouping variables, matching variables, groups to be considered, binary or multigroup analyses, which group is termed the reference group, IRT or no-IRT approaches, how much DIF is significant, what to do with items that function differentially, etc., posit DIF analyses more subjective than meets the eye.

Some of these decisions are made on an ontological level, where the assumption is that a universal construct exists on which all examinees have a trustworthy standing. This ontology informs the questions we ask. Others are made on epistemological levels, where the social history of White Euro-Americans determines what counts as evidence and what ways of seeking it are legitimate (Scheurich & Young, 1997). These epistemologies have a bearing on how researchers implement methods. Some decisions are based on the constraints of the study, such as how much data is available. Others are made by the researcher as a person and are informed by their beliefs and experiences.

It follows that if researchers using DIF analyses make all these subjective decisions in their studies, I also make subjective decisions in this study informed by my personal history, identity, philosophical and political beliefs, ontologies, and epistemologies. As a Black person, I am largely on the fence, trained formally in Western epistemologies, ontologies and axiologies yet acutely aware of the alternate, albeit marginalized, epistemologies, ontologies and axiologies that characterize my African culture.

There are several salient differences in the experience of being Black in Africa and being Black in the USA. The biggest is the notion of which racial group is in the majority and which is in the minority and the effect that has on how the terminology is used. In my upbringing, the racial group in the majority in terms of population statistics was Black; thus, the term majority, in my mind, referred to the numerical majority. In the USA, on the other hand, the majority refers to the numerical majority. However, the term holds even in local contexts, such that one can teach in a majority-minority school, which means that there are mostly minority students in that school.

There are several reasons why this section is important for this study. Firstly, the research study explored race and ethnicity as variables in research. My lived experience as a Black international female student living in the USA surfaced and resurfaced many questions around identity in general and my identity in particular. Secondly, the knowledge generated by research studies, including those in my study sample, bears on numerous aspects of my life. I found myself, at all points in the research journey, frequently called to sit and struggle with the interaction of myself, my beliefs, past experiences, current experiences, hopes and wishes for the future with my research study, the role of race and ethnicity in research and the field of psychometrics.

My personal and professional background in Chapter 3 illustrates why I was drawn to this research study. I believe there is no universal psychological construct that is unchanged across different cultures. I believe that what is valued in one culture or tradition is not necessarily so in another and that while neither is better, each is a potentially good fit for its context. To explain, I believe that the numeracy skills fundamental to survival in a US city differ from those needed in peri-urban Lesotho. As such, I believe that constructs, instruments, and studies need better identification in terms of the cultures and contexts that created them such that an internationally administered mathematics proficiency is described in terms of who was present when the construct was specified, the backgrounds of those who write items, the repositories of examples they consulted, and their values. The following section describes how my beliefs and values and the research study interfaced and what I did about it.

A review of the open-ended comments on the data quantification form shows that early in the quantification process, research designs deviated from the traditional use of DIF for educational or psychological tests. Comments were also mainly related to the definitions of race

and ethnicity provided. These early reflections became prompts for engaging peers and faculty on research designs and the applicability of DIF analyses to general research. As quantification progressed, open-ended comments were related less to methodology and more to the treatment of race and ethnicity either in the methodology section or as findings were presented and interpreted, which were incorporated into Chapter IV to highlight special and non-standard cases.

Reflective conversations with faculty and peers during data quantification also served as a quality assurance measure, considering the limitation of only one researcher described earlier. The format of the reflective conversions was a review of the abstract and methods section of an article followed by a discussion guided by pointed questions. Conversations also included a review of the definition of DIF, specifically the creation of groups, the person whose race and ethnicity is collected, reported and used as a grouping variable (e.g., in studies where the respondent provided scores on behalf of or in relations to someone else such a child, student or patient), and the matching variable used. Discussions were bolstered by formal definitions and, in some cases, the expertise and experience with DIF of faculty members, and when consensus was reached, it informed the refinement of coding/quantification protocols.

My reflective journal also revealed a progression in thoughts as the study progressed. In the beginning, reflections were restricted to the methodology employed by the articles being quantified/sampled, especially when they were novel or unfamiliar. I noticed that reflective entries tended to be longer for studies that presented no definition or explicit conceptualization of race and ethnicity, especially when findings rendered race or ethnicity as causal variables. I also noticed that my attention was caught by those articles that minimized the chances of the observed DIF being caused by biased items based on the minimal impact of DIF. I reflected on

my research questions and rationalized the quantification variables to remove the distraction caused by the presence/absence and the nature of follow-up bias studies.

I also found myself drawn to the few research articles with explicit definitions/conceptualizations of race and ethnicity and those that brought the historical and political context into the discussion of findings. I found that these characterizations.

For example, a guest speaker at a student-organized conference in the ERM department was an evaluator/researcher of American Indian descent who, in passing, discussed the consistent erasure of the American Indian racial group in contemporary research, methodology and theorizing. As an international student, and perhaps because my focus was on how my racial group is treated in research studies, I realized then that I had also been complicit in this erasure as my data quantification form did not even include American Indians as one of the racial/ethnic groups. I adjusted the form following the guest lecture.

Limitations

This section discusses the limitations of the present study. The limitations are presented in two broad categories: those related to Phase 1 and Phase 2. In the first category are issues related to the literature search, including over-representation of certain projects (e.g., PROMIS) and authors, limited accessibility to databases used in previous systematic literature reviews, and the fact that I was the only researcher in the study, which is not ideal according to PRISMA standards. In the second category, limitations were encountered in Phase 2 of the study, including challenges with the recruitment of participants and the implementation of the reflection on forward citations.

Phase 1

A big limitation of the systematic literature review is that I was the only person reviewing and quantifying the research articles. The recommendation for reviews of this scope is to have two or more reviewers (Grant & Booth, 2009). I enlisted the help of a colleague in the screening of articles for inclusion in Phase 1. However, for the actual data extraction and analysis, it was not possible to have a second reviewer.

Another limitation of Phase 1 was the availability of databases provided by the UNCG Library. Previous systematic literature reviews related to DIF (Berrío et al., 2020; Gómez Benito et al., 2005) used the Web of Science online database, which was unavailable through the UNCG Library. As such, there is a big possibility that eligible articles were excluded from the review due to location bias. In addition, DIF studies conducted in the Education sector are conducted by testing companies and are not always published in peer-review journals; thus, those DIF studies were not available for selection.

Another limitation of Phase 1 was the overrepresentation of articles from the Patient Reported Outcomes Measurement Information System (PROMIS) program, which aimed to develop and validate various measures and question banks related to physical, mental and social health in adults and children. DIF analyses formed a big part of the program, and all 18 articles related to PROMIS measures and item banks had similar study procedures. Additionally, most of the 18 articles related to PROMIS were published in special journal issues.

Phase 2

The goal of the study design was to interview 12 authors from Phase 1, so the low response rate and ultimate sample size of the two articles presented some limitations. Both articles included in Phase 2 conducted DIF analyses on tools related to psychological constructs.

Both authors interviewed were older white males who were late-career researchers in their respective psychology subfields. While there were differences in the conceptualization of race and ethnicity in the two articles sampled for Phase 2, the lack of diversity in researchers was a limitation, as it is challenging to understand the full range of perspectives. In addition, with the limited response rate for Phase 2 participants, it is difficult to attain data saturation.

Another limitation experienced in Phase 2 was the limited response to the forward citations. The authors were provided with the forward citations prior to their interview, but they did not read the forward citations; thus, their reflections were limited to what they were able to glean in the second half of their interview.

REFERENCES

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using Multidimensional Item Response Theory to Evaluate Educational and Psychological Tests. *Educational Measurement: Issues and Practice*, *22*(3), 37–51. https://doi.org/10.1111/j.1745-3992.2003.tb00136.x

Adams, D., Joo, M. T. H., Sumintono, B., & Pei, O. S. (2020). Blended learning engagement in higher education institutions: A differential item functioning analysis of students' backgrounds. *Malaysian Journal of Learning and Instruction*, *17*(Number 1), 133–158. https://doi.org/10.32890/mjli2020.17.1.6

Aiken, L. R. (1985). *Psychological testing and assessment* (5th ed). Allyn and Bacon.

Aksu-Dunya, B., McKown, C., & Smith, E. (2020). Psychometric Properties and Differential Item Functioning of a Web-Based Assessment of Children's Emotion Recognition Skill. *Journal of Psychoeducational Assessment*, *38*(5), 627–641. https://doi.org/10.1177/0734282919881919

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.

American Psychological Association. (2023). *APA Dictionary of Psychology*. https://dictionary.apa.org/

Anastasi, A. (1976). *Psychological testing* (4th ed). Macmillan.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In *Differential item functioning.* (pp. 3–23). Lawrence Erlbaum Associates, Inc. https://doi.org/10.1075/z.62.13kok

Armenta, B. E., & Cooper, M. L. (2019). The Rutgers Alcohol Problem Index: Measurement equivalence among college students in the US and Mexico. *Psychological Assessment, 31*(1), 1–14. https://doi.org/10.1037/pas0000608

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., Fiedler, S., Funder, D. C., Kliegl, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., Van Aken, M. A. G., Weber, H., & Wicherts, J. M. (2013). Recommendations for Increasing Replicability in Psychology. *European Journal of Personality*, *27*(2), 108–119. https://doi.org/10.1002/per.1919

Baker, Dominique J., Ford, Karly S., Viano, Samantha, & Johnston-Guerrero, Marc P. (2022). Racial Category Usage in Education Research: Examining the Publications from AERA Journals. *EdWorkingPaper: 22-596*. https://doi.org/10.26300/R9DG-KD13

Bandalos, D. (2018). 16. Bias, Fairness, and Legal Issues. In *Measurement Theory and Applications for the Social Sciences*. Guildford Press.

Benítez, I., Padilla, J.-L., Hidalgo Montesinos, M. D., & Sireci, S. G. (2016). Using Mixed Methods to Interpret Differential Item Functioning. *Applied Measurement in Education*, *29*(1), 1–16. https://doi.org/10.1080/08957347.2015.1102915

Berrío, Á. I., Gómez-Benito, J., & Arias-Patiño, E. M. (2020). Developments and trends in research on methods of detecting differential item functioning. *Educational Research Review*, *31*, 100340. https://doi.org/10.1016/j.edurev.2020.100340

Bowe, A. (2017). The cultural fairness of the 12-item General Health Questionnaire among diverse adolescents. *Psychological Assessment*, *29*(1), 87–97. https://doi.org/10.1037/pas0000323

Bowe, A. (2019). Moving Toward More Conclusive Measures of Sociocultural Adaptation for Ethnically Diverse Adolescents in England. *Canadian Journal of School Psychology*, *34*(1), 56–72. https://doi.org/10.1177/0829573517739392

Brown, K., & Jackson, D. D. (2014). The History and Conceptual Elements of Critical Race Theory. In *Handbook of Critical Race Theory in Education*. Routledge. https://doi.org/10.4324/9780203155721.ch1

Camilli, G. (1992). A Conceptual Analysis of Differential Item Functioning in Terms of a Multidimensional Item Response Model. *Applied Psychological Measurement*, *16*(2), 129–147. https://doi.org/10.1177/014662169201600203

Camilli, G., & Shepard, L. A. (2022). *Methods for Identifying Biased Test Items* (Vol. 4). SAGE Publications, Inc. https://us.sagepub.com/en-us/nam/methods-for-identifying-biased-test-items/book3416

Castagno, A. E. (2005). Extending the Bounds of Race and Racism: Indigenous Women and the Persistence of the Black–White Paradigm of Race. *The Urban Review*, *37*(5), 447–468. https://doi.org/10.1007/s11256-005-0020-4

Chakawa, A., Butler, R. C., & Shapiro, S. K. (2015). Examining the psychometric validity of the Multigroup Ethnic Identity Measure-Revised (MEIM-R) in a community sample of African American and European American adults. *Cultural Diversity and Ethnic Minority Psychology*, *21*(4), 643–648. https://doi.org/10.1037/cdp0000025

Chen, H. F., & Zhu, J. (2017). Optimal items for assessing parental involvement across different groups during middle childhood. *Journal of Child and Family Studies*, *26*(11), 2999–3012. https://doi.org/10.1007/s10826-017-0809-2

Chen, T.-A., O'Connor, T. M., Hughes, S. O., Beltran, A., Baranowski, J., Diep, C., & Baranowski, T. (2015). Vegetable parenting practices scale. Item response modeling analyses. *Appetite*, *91*, 190–199. https://doi.org/10.1016/j.appet.2015.04.048

Clauser, B. E., & Mazor, K. M. (1998). Using Statistical Procedures to Identify Differentially Functioning Test Items. *Educational Measurement: Issues and Practice*, *17*(1), 31–44. https://doi.org/10.1111/j.1745-3992.1998.tb00619.x

Cole, N. S. (1993). History and Development of DIF. In *Differential Item Functioning* (pp. 25–34). Lawrence Erlbaum Associates, Inc.

Cordier, R., Munro, N., Wilkes-Gillan, S., Speyer, R., Parsons, L., & Joosten, A. (2019). Applying Item Response Theory (IRT) Modeling to an Observational Measure of Childhood Pragmatics: The Pragmatics Observational Measure-2. *Frontiers in Psychology*, *10*, 408. https://doi.org/10.3389/fpsyg.2019.00408

Covarrubias, A., Nava, P. E., Lara, A., Burciaga, R., & Solórzano, D. G. (2018). Expanding Educational Pipelines. In J. T. DeCuir-Gunby, T. K. Chapman, & P. A. Schutz (Eds.), *Understanding Critical Race Research Methods and Methodologies* (1st ed., pp. 138–149). Routledge. https://doi.org/10.4324/9781315100944-12

Covarrubias, A., & Vélez, V. N. (2013). Critical Race Quantitative Intersectionality: An Anti-Racist Research Paradigm that Refuses to "Let the Numbers Speak for Themselves." In M. Lynn & A. D. Dixson (Eds.), *Handbook of Critical Race Theory in Education* (0 ed., pp. 290–306). Routledge. https://doi.org/10.4324/9780203155721-30

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed). Pearson.

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (Third Edition). SAGE.

Davis, J. E. (1992). Reconsidering the use of race as an explanatory variable in program evaluation. *New Directions for Program Evaluation*, *1992*(53), 55–67. https://doi.org/10.1002/ev.1601

de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. The Guilford Press.

Delgado, R., & Stefancic, J. (2021). Discerning Critical Moments. In M. Lynn & A. D. Dixson, *Handbook of Critical Race Theory in Education* (2nd ed., pp. 22–31). Routledge. https://doi.org/10.4324/9781351032223-4

Delgado, R., & Stefancic, J. (2017). *Critical race theory: An introduction* (Third edition). New York University Press.

Devine, R. T., & Hughes, C. (2016). Measuring theory of mind across middle childhood: Reliability and validity of the Silent Films and Strange Stories tasks. *Journal of Experimental Child Psychology*, *149*, 23–40. https://doi.org/10.1016/j.jecp.2015.07.011

Dixon-Román, E. (2020). A Haunting Logic of Psychometrics: Toward the Speculative and Indeterminacy of Blackness in Measurement. *Educational Measurement: Issues and Practice*, *39*(3), 94–96. https://doi.org/10.1111/emip.12375

Dixson, A. D., & Rousseau, C. K. (2005). And we are still not saved: Critical race theory in education ten years later. *Race Ethnicity and Education*, *8*(1), 7–27. https://doi.org/10.1080/1361332052000340971

Dmitrieva, N. O., Fyffe, D., Mukherjee, S., Fieo, R., Zahodne, L. B., Hamilton, J., Potter, G. G., Manly, J. J., Romero, H. R., Mungas, D., & Gibbons, L. E. (2015). Demographic characteristics do not decrease the utility of depressive symptoms assessments:

Examining the practical impact of item bias in four heterogeneous samples of older adults: Differential item function in depressive symptoms. *International Journal of Geriatric Psychiatry*, *30*(1), 88–96. https://doi.org/10.1002/gps.4121

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In *Differential item functioning.* (pp. 35–66). Lawrence Erlbaum Associates, Inc. https://doi.org/10.1075/z.62.13kok

Du Plessis, G. A., & De Bruin, G. P. (2015). Using Rasch modelling to examine the international personality item pool (IPIP) values in action (VIA) measure of character strengths. *Journal of Psychology in Africa*, *25*(6), 512–521. https://doi.org/10.1080/14330237.2015.1124603

DuPaul, G. J., Fu, Q., Anastopoulos, A. D., Reid, R., & Power, T. J. (2020). ADHD parent and teacher symptom ratings: Differential item functioning across gender, age, race, and ethnicity. *Journal of Abnormal Child Psychology*, *48*(5), 679–691. https://doi.org/10.1007/s10802-020-00618-7

Ehrich, J., Howard, S. J., Mu, C., & Bokosmaty, S. (2016). A comparison of Chinese and Australian university students' attitudes towards plagiarism. *Studies in Higher Education*, *41*(2), 231–246. https://doi.org/10.1080/03075079.2014.927850

Eugenics and Scientific Racism. (n.d.). Genome.Gov. Retrieved January 14, 2023, from https://www.genome.gov/about-genomics/fact-sheets/Eugenics-and-Scientific-Racism

Farrington, A. L., & Lonigan, C. J. (2015). Examining the Measurement Precision and Invariance of the Revised Get Ready to Read! *Journal of Learning Disabilities*, *48*(3), 227–238. https://doi.org/10.1177/0022219413495568

Fieo, R., Ocepek-Welikson, K., Kleinman, M., Eimicke, P., Crane, P. K., Cella, D., & Teresi, J. A. (2016). Measurement equivalence of the Patient Reported Outcomes Measurement Information System® (PROMIS®) Applied Cognition – General Concerns, short forms in ethnically diverse groups. *Psychological Test and Assessment Modeling*, *58*(2), 255–307.

Filshtein, T., Chan, M., Mungas, D., Whitmer, R., Fletcher, E., DeCarli, C., & Farias, S. (2020). Differential Item Functioning of the Everyday Cognition (ECog) Scales in Relation to Racial/Ethnic Groups. *Journal of the International Neuropsychological Society*, *26*(5), 515–526. https://doi.org/10.1017/S1355617719001437

Finch, H. (2005). The MIMIC Model as a Method for Detecting DIF: Comparison With Mantel-Haenszel, SIBTEST, and the IRT Likelihood Ratio. *Applied Psychological Measurement*, *29*(4), 278–295. https://doi.org/10.1177/0146621605275728

Flanagan, A. M., Cormier, D. C., & Bulut, O. (2020). Achievement may be rooted in teacher expectations: Examining the differential influences of ethnicity, years of teaching, and classroom behaviour. *Social Psychology of Education: An International Journal*, *23*(6), 1429–1448. https://doi.org/10.1007/s11218-020-09590-y

Forrest, C. B., Devine, J., Bevans, K. B., Becker, B. D., Carle, A. C., Teneralli, R. E., Moon, J., Tucker, C. A., & Ravens-Sieberer, U. (2018). Development and psychometric evaluation of the PROMIS Pediatric Life Satisfaction item banks, child-report, and parent-proxy editions. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation*, *27*(1), 217–234. https://doi.org/10.1007/s11136-017-1681-7

Forrest, C. B., Ravens-Sieberer, U., Devine, J., Becker, B. D., Teneralli, R. E., Moon, J., Carle, A. C., Tucker, C. A., & Bevans, K. B. (2018). Development and evaluation of the PROMIS® Pediatric Positive Affect item bank, child-report and parent-proxy editions. *Journal of Happiness Studies: An Interdisciplinary Forum on Subjective Well-Being*, *19*(3), 699–718. https://doi.org/10.1007/s10902-016-9843-9

Galenkamp, H., Stronks, K., Mokkink, L. B., & Derks, E. M. (2018). Measurement invariance of the SF-12 among different demographic groups: The HELIUS study. *PLoS ONE*, *13*(9). https://doi.org/10.1371/journal.pone.0203483

Galenkamp, H., Stronks, K., Snijder, M. B., & Derks, E. M. (2017). Measurement invariance testing of the PHQ-9 in a multi-ethnic population in Europe: The HELIUS study. *BMC Psychiatry*, *17*. https://doi.org/10.1186/s12888-017-1506-9

Garcia, N. M., López, N., & Vélez, V. N. (2018). QuantCrit: Rectifying quantitative methods through critical race theory. *Race Ethnicity and Education*, *21*(2), 149–157. https://doi.org/10.1080/13613324.2017.1377675

Gay, C. L., Kottorp, A., Lerdal, A., & Lee, K. A. (2016). Psychometric limitations of the Center for Epidemiologic Studies-Depression scale for assessing depressive symptoms among adults with HIV/AIDS: A Rasch analysis. *Depression Research and Treatment*, *2016*. https://doi.org/10.1155/2016/2824595

Geldenhuys, M., & Bosch, A. (2020). A Rasch adapted version of the 30-item Bem Sex Role Inventory (BSRI). *Journal of Personality Assessment*, *102*(3), 428–439. https://doi.org/10.1080/00223891.2018.1527343

Gierl, M. J. (2005). Using Dimensionality-Based DIF Analyses to Identify and Interpret Constructs That Elicit Group Differences. *Educational Measurement: Issues and Practice*, *24*(1), 3–14. https://doi.org/10.1111/j.1745-3992.2005.00002.x

Gillborn, D., Warmington, P., & Demack, S. (2018). QuantCrit: Education, policy, 'Big Data' and principles for a critical race theory of statistics. *Race Ethnicity and Education*, *21*(2), 158–179. https://doi.org/10.1080/13613324.2017.1377417

Goetz, C. G., Liu, Y., Stebbins, G. T., Wang, L., Tilley, B. C., Teresi, J. A., Merkitch, D., & Luo, S. (2016). Gender-, age-, and race/ethnicity-based differential item functioning analysis of the movement disorder society-sponsored revision of the Unified Parkinson's disease rating scale: DIF Analysis Of MDS-UPDRS. *Movement Disorders*, *31*(12), 1865–1873. https://doi.org/10.1002/mds.26847

Gómez-Benito, J., Sireci, S., Padilla, J.-L., Hidalgo, M. D., & Benítez, I. (2018). Differential Item Functioning: Beyond validity evidence based on internal structure: Funcionamiento Diferencial del Item: más allá de las evidencias de validez basadas en la estructura interna. *Psicothema*, *30*(1), 104–109. https://doi.org/10.7334/psicothema2017.183

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal*, *26*(2), 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

Hahn, E. A., Kallen, M. A., Jensen, R. E., Potosky, A. L., Moinpour, C. M., Ramirez, M., Cella, D., & Teresi, J. A. (2016). Measuring social function in diverse cancer populations: Evaluation of measurement equivalence of the Patient Reported Outcomes Measurement Information System® (PROMIS®) Ability to Participate in Social Roles and Activities short form. *Psychological Test and Assessment Modeling*, *58*(2), 403–421.

Haroz, E. E., Bolton, P., Gross, A., Chan, K. S., Michalopoulos, L., & Bass, J. (2016). Depression symptoms across cultures: An IRT analysis of standard depression symptoms using data from eight countries. *Social Psychiatry and Psychiatric Epidemiology: The International Journal for Research in Social and Genetic Epidemiology and Mental Health Services*, *51*(7), 981–991. https://doi.org/10.1007/s00127-016-1218-3

Harpole, J. K., Levinson, C. A., Woods, C. M., Rodebaugh, T. L., Weeks, J. W., Brown, P. J., Heimberg, R. G., Menatti, A. R., Blanco, C., Schneier, F., & Liebowitz, M. (2015). Assessing the straightforwardly-worded Brief Fear of Negative Evaluation Scale for differential item functioning across gender and ethnicity. *Journal of Psychopathology and Behavioral Assessment*, *2*, 306–317. https://doi.org/10.1007/s10862-014-9455-9

Hasnain, M., Gruss, V., Keehn, M., Peterson, E., Valenta, A. L., & Kottorp, A. (2017). Development and validation of a tool to assess self-efficacy for competence in interprofessional collaborative practice. *Journal of Interprofessional Care*, *31*(2), 255–262. https://doi.org/10.1080/13561820.2016.1249789

Hawes, S. W., Byrd, A. L., Kelley, S. E., Gonzalez, R., Edens, J. F., & Pardini, D. A. (2018). Psychopathic features across development: Assessing longitudinal invariance among Caucasian and African American youths. *Journal of Research in Personality*, *73*, 180–188. https://doi.org/10.1016/j.jrp.2018.02.003

Heafner, T. L., & Fitchett, P. G. (2018). US history content knowledge and associated effects of race, gender, wealth, and urbanity: Item Response Theory (IRT) modeling of NAEP-USH achievement. *Journal of Social Studies Research*, *42*(1), 11–25. https://doi.org/10.1016/j.jssr.2017.01.001

Healy, L. M. (2006). *Running Head: LOGISTIC REGRESSION: AN OVERVIEW Logistic Regression: An Overview*.

Holland, P. W. (2008). Causation and race. In T. Zuberi & E. Bonilla-Silva (Eds.), *White logic, white methods: Racism and methodology*. Rowman & Littlefield Publishers.

Holland, W., & Thayer, D. (1998). "Univariate and Bivariate Loglinear Models for Discrete Test Score Distributions." ETS Research Report Series, vol. 1998, no. 2, 1998, pp. i–56, doi:10.1002/j.2333-8504.1998.tb01776.x.

Hong, I., Velozo, C. A., Li, C.-Y., Romero, S., Gruber-Baldini, A. L., & Shulman, L. M. (2016). Assessment of the psychometrics of a PROMIS item bank: Self-efficacy for managing daily activities. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation*, *25*(9), 2221–2232. https://doi.org/10.1007/s11136-016-1270-1

Humes, K., Jones, N.A. and Ramirez, R.R. (2011) Overview of Race and Hispanic Origin: 2010. U.S. Census Bureau. http://www.census.gov/prod/cen2010/briefs/c2010br-02.pdf

Hu, J., Serovich, J., Brown, M., Kimberly, J., & Chen, Y.-H. (2017). Psychometric Evaluation of the HIV Disclosure Belief Scale: A Rasch Model Approach. *AIDS & Behavior*, *21*(1), 174–183. https://doi.org/10.1007/s10461-016-1478-7

Jackson, J. P., & Weidman, N. M. (2004). *Race, racism, and science: Social impact and interaction*. ABC-CLIO.

James, A. (2008). Making sense of race and racial classification. In T. Zuberi & E. Bonilla-Silva (Eds.), *White logic, white methods: Racism and methodology*. Rowman & Littlefield Publishers.

Janulis, P., Newcomb, M. E., Sullivan, P., & Mustanski, B. (2018). Evaluating HIV knowledge questionnaires among men who have sex with men: A multi-study item response theory analysis. *Archives of Sexual Behavior*, *47*(1), 107–119. https://doi.org/10.1007/s10508-016-0910-4

Jensen, R. E., King-Kallimanis, B. L., Sexton, E., Reeve, B. B., Moinpour, C. M., Potosky, A. L., Lobo, T., & Teresi, J. A. (2016). Measurement properties of PROMIS Sleep Disturbance short forms in a large, ethnically diverse cancer cohort. *Psychological Test and Assessment Modeling*, *58*(2).

Jones, L. V., & Thissen, D. (2007). A History and Overview of Psychometrics. *Handbook of Statistics*, *26*, 1–27. https://doi.org/10.1016/S0169-7161(06)26001-2

Jonson, J. L., & Geisinger, K. F. (Eds.). (2022). *Fairness in educational and psychological testing: Examining theoretical, research, practice, and policy implications of the 2014 standards*. American Educational Research Association.

Kane, M. (2006). Content-Related Validity Evidence in Test Development. In *Handbook of test development.* (pp. 131–153). Lawrence Erlbaum Associates Publishers.

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores: Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kaufman, J. S., Cooper, R. S., & McGee, D. L. (1997). Socioeconomic status and health in Blacks and Whites: The problem of residual confounding and the resiliency of race. Epidemiology (Cambridge, Mass.), 8(6), 621–628.

Khalfani, A. K., Zuberi, T., Bah, S., & Lehohla, P. J. (2008). Race and population statistics in South Africa. In T. Zuberi & E. Bonilla-Silva (Eds.), *White logic, white methods: Racism and methodology*. Rowman & Littlefield Publishers.

Kim, G., DeCoster, J., Bryant, A. N., & Ford, K. L. (2016). Measurement Equivalence of the K6 Scale. *Assessment*, *23*(6), 758–768. https://doi.org/10.1177/1073191115599639

Kivisto, P., & Croll, P. R. (2012). *Race and ethnicity: The basics*. Routledge.

Kwan, Y. H., Uy, E. J., Bautista, D. C., Xin, X., Xiao, Y., Lee, G. L., Subramaniam, M., Vaingankar, J. A., Chan, M. F., Kumar, N., Cheung, Y. B., Chua, T. S. J., & Thumboo, J. (2019a). Development and calibration of a novel positive mindset item bank to measure health-related quality of life (HRQoL) in Singapore. *PLoS ONE*, *14*(7). https://doi.org/10.1371/journal.pone.0220293

Kwan, Y. H., Uy, E. J., Bautista, D. C., Xin, X., Xiao, Y., Lee, G. L., Subramaniam, M., Vaingankar, J. A., Chan, M. F., Kumar, N., Cheung, Y. B., Chua, T. S. J., & Thumboo, J. (2019b). Development and calibration of a novel social relationship item bank to measure health-related quality of life (HRQoL) in Singapore. *Health and Quality of Life Outcomes*, *17*. https://doi.org/10.1186/s12955-019-1150-9

Ladson-Billings, G., & Tate, W. F. (1995). *Toward a Critical Race Theory of Education*. 22.

Lambert, M. C., Garcia, A. G., Epstein, M. H., & Cullinan, D. (2018). Differential Item Functioning of the Emotional and Behavioral Screener for Caucasian and African American Elementary School Students. *Journal of Applied School Psychology*, *34*(3), 201–214. https://doi.org/10.1080/15377903.2017.1345815

Lambert, M. C., January, S.-A. A., Cress, C. J., Epstein, M. H., & Cullinan, D. (2018). Differential Item Functioning across Race and Ethnicity for the Emotional and Behavioral Screener. *School Psychology Quarterly*, *33*(3), 399–407.

Lane, S. (1999). *Validity Evidence for Assessments*. 20.

Lange, A. M. C., Scholte, R. H. J., van Geffen, W., Timman, R., Busschbach, J. J. V., & van der Rijken, R. E. A. (2016). The lack of cross-national equivalence of a Therapist Adherence Measure (TAM-R) in multisystemic therapy (MST). *European Journal of Psychological Assessment*, *32*(4), 312–325. https://doi.org/10.1027/1015-5759/a000262

Lindhiem, O., Vaughn-Coaxum, R. A., Higa, J., Harris, J. L., Kolko, D. J., & Pilkonis, P. A. (2019). Development and validation of the Knowledge of Effective Parenting Test (KEPT) in a nationally representative sample. *Psychological Assessment*, *31*(6), 781–792. https://doi.org/10.1037/pas0000699

Loyd, A. B., Derlan, C. L., Smith, E. V., Norris, S. A., Richter, L. M., & Roeser, R. W. (2019). Evaluating the Psychometric Properties of a Measure of Ethnic Identity Among Black South African Youth. *Identity*, *19*(1), 44–61. https://doi.org/10.1080/15283488.2019.1566070

Manly, J., T. (2006). Deconstructing Race and Ethnicity: Implications for Measurement of Health Outcomes. *Medical Care*, *44*(1).

Marks, A & García C, C. (2012). The Immigrant Paradox in Children and Adolescents: Is Becoming American a Developmental Risk?. 10.1037/13094-000.

Martin, J. L., Colvin, K. F., Madson, M. B., Zamboanga, B. L., & Pazienza, R. (2020). Optimal assessment of protective behavioral strategies among college drinkers: An item response

theory analysis. *Psychological Assessment*, *32*(4), 394–406. https://doi.org/10.1037/pas0000799

Mateos, P., Singleton, A., & Longley, P. (2009). Uncertainty in the Analysis of Ethnicity Classifications: Issues of Extent and Aggregation of Ethnic Groups. *Journal of Ethnic and Migration Studies*, *35*(9), 1437–1460. https://doi.org/10.1080/13691830903125919

Messick, S. (1989). Validity. In *Educational measurement, 3rd ed.* (pp. 13–103). American Council on Education.

Miller, A. P., Merkle, E. C., Galenkamp, H., Stronks, K., Derks, E. M., & Gizer, I. R. (2019). Differential item functioning analysis of the CUDIT and relations with alcohol and tobacco use among men across five ethnic groups: The HELIUS study. *Psychology of Addictive Behaviors*, *33*(8), 697–709. https://doi.org/10.1037/adb0000521

National Institutes of Health. (2015). *NOT-OD-15-089: Racial and Ethnic Categories and Definitions for NIH Diversity Programs and for Other Reporting Purposes*. https://grants.nih.gov/grants/guide/notice-files/NOT-OD-15-089.html

Ning, H. K. (2018). A Rasch analysis of the Junior Metacognitive Awareness Inventory with Singapore students. *Measurement & Evaluation in Counseling & Development*, *51*(2), 84–91. https://doi.org/10.1080/07481756.2017.1358061

Office of Management and Budget. (1997). *Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity*. https://obamawhitehouse.archives.gov/node/15626

Osterlind, S. J., & Everson, H. T. (2009). *Differential Item Functioning* (2nd Edition). SAGE Publishing.

Park, I. H., Rachmatullah, A., Park, I.-S., & Liu, Y. (2019). Assessment of the quality and generalizability of the revised UCLA loneliness scale in Chinese and Korean

community-dwelling elderly populations using item response theory (IRT)-Rasch modeling and hybrid IRT-logistic regression. *Educational Gerontology*, *45*(10), 581–599. https://doi.org/10.1080/03601277.2019.1670908

Parkerson, H. A., Thibodeau, M. A., Brandt, C. P., Zvolensky, M. J., & Asmundson, G. J. G. (2015). Cultural-based biases of the GAD-7. *Journal of Anxiety Disorders*, *31*, 38–42. https://doi.org/10.1016/j.janxdis.2015.01.005

Pedersen, E. R., Huang, W., Dvorak, R. D., Prince, M. A., & Hummer, J. F. (2017). The Protective Behavioral Strategies for Marijuana Scale: Further examination using item response theory. *Psychology of Addictive Behaviors*, *31*(5), 548–559. https://doi.org/10.1037/adb0000271

Peipert, J. D., Bentler, P., Klicko, K., & Hays, R. D. (2018). Negligible impact of differential item functioning between Black and White dialysis patients on the Kidney Disease Quality of Life 36-item short form survey (KDQOLTM-36). *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation*, *27*(10), 2699–2707. https://doi.org/10.1007/s11136-018-1879-3

Plake, B. S., & Wise, L. L. (2014). What Is the Role and Importance of the Revised AERA, APA, NCME Standards for Educational and Psychological Testing? *Educational Measurement: Issues and Practice*, *33*(4), 4–12. https://doi.org/10.1111/emip.12045

Poe, M. (2009). Reporting race and ethnicity in international assessment. In C. S. Schreiner (Ed.), *Handbook of research on assessment technologies, methods, and applications in higher education* (p. 500). 10.4018/978-1-60566-667-9

Quach, C. W., Langer, M. M., Chen, R. C., Thissen, D., Usinger, D. S., Emerson, M. A., & Reeve, B. B. (2016). Reliability and validity of PROMIS measures administered by

telephone interview in a longitudinal localized prostate cancer study. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation*, *25*(11), 2811–2823. https://doi.org/10.1007/s11136-016-1325-3

Reeve, B. B., Pinheiro, L. C., Jensen, R. E., Teresi, J. A., Potosky, A. L., McFatrich, M. K., Ramirez, M., & Chen, W.-H. (2016). Psychometric evaluation of the PROMIS® Fatigue measure in an ethnically and racially diverse population-based sample of cancer patients. *Psychological Test and Assessment Modeling*, *58*(1), 119–139.

Reich, H., Rief, W., Brähler, E., & Mewes, R. (2018). Cross-cultural validation of the German and Turkish versions of the PHQ-9: An IRT approach. *BMC Psychology*, *6*(1), 26. https://doi.org/10.1186/s40359-018-0238-z

Rice, S. M., Parker, A. G., Mawren, D., Clifton, P., Harcourt, P., Lloyd, M., Kountouris, A., Smith, B., McGorry, P. D., & Purcell, R. (2020). Preliminary psychometric validation of a brief screening tool for athlete mental health among male elite athletes: The Athlete Psychological Strain Questionnaire. *International Journal of Sport and Exercise Psychology*, *18*(6), 850–865. https://doi.org/10.1080/1612197X.2019.1611900

Rose, A. J., Bayliss, E., Huang, W., Baseman, L., Butcher, E., García, R.-E., & Edelen, M. O. (2018). Evaluating the PROMIS-29 v20 for use among older adults with multiple chronic conditions. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation*, *27*(11), 2935–2944. https://doi.org/10.1007/s11136-018-1958-5

Ross, P. T., Hart-Johnson, T., Santen, S. A., & Zaidi, N. L. B. (2020). Considerations for using race and ethnicity as quantitative variables in medical education research. *Perspectives on Medical Education*, *9*(5), 318–323. https://doi.org/10.1007/s40037-020-00602-3

Roth, W. D. (2016). The multiple dimensions of race. *Ethnic and Racial Studies*, *39*(8), 1398–1406. https://doi.org/10.1080/01419870.2016.1140793

Roussos, L., & Stout, W. (1996). A Multidimensionality-Based DIF Analysis Paradigm. *Applied Psychological Measurement*, *20*(4), 355–371. https://doi.org/10.1177/014662169602000404

Roussos, L., & Stout, W. (2004). Differential Item Functioning Analysis: Detecting DIF Items and Testing DIF Hypotheses. In D. Kaplan, *The SAGE Handbook of Quantitative Methodology for the Social Sciences* (pp. 108–117). SAGE Publications, Inc. https://doi.org/10.4135/9781412986311.n6

Roy, C., Bakan, G., Li, Z., & Nguyen, T. H. (2016). Coping measurement: Creating short form of Coping and Adaptation Processing Scale using item response theory and patients dealing with chronic and acute health conditions. *Applied Nursing Research*, *32*, 73–79. https://doi.org/10.1016/j.apnr.2016.06.002

Russell, M. K. (2024). *Systemic racism and educational measurement: Confronting injustice in testing, assessment, and beyond* (First edition). Routledge.

Rust, J., Kosinski, M., & Stillwell, D. (2020). *Modern Psychometrics: The Science of Psychological Assessment* (4th ed.). Routledge. https://doi.org/10.4324/9781315637686

Sablan, J. R. (2019). Can You Really Measure That? Combining Critical Race Theory and Quantitative Methods. *American Educational Research Journal*, *56*(1), 178–203. https://doi.org/10.3102/0002831218798325

Salsman, J. M., Schalet, B. D., Merluzzi, T. V., Park, C. L., Hahn, E. A., Snyder, M. A., & Cella, D. (2019). Calibration and initial validation of a general self-efficacy item bank and short form for the NIH PROMIS®. *Quality of Life Research*, *28*(9), 2513–2523. https://doi.org/10.1007/s11136-019-02198-6

Salsman, J. M., Schalet, B. D., Park, C. L., George, L., Steger, M. F., Hahn, E. A., Snyder, M. A., & Cella, D. (2020). Assessing meaning & purpose in life: Development and validation of an item bank and short forms for the NIH PROMIS®. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation*, *29*(8), 2299–2310. https://doi.org/10.1007/s11136-020-02489-3

Sandham, M. H., Medvedev, O. N., Hedgecock, E., Higginson, I. J., & Siegert, R. J. (2019). A Rasch Analysis of the Integrated Palliative Care Outcome Scale. *Journal of Pain and Symptom Management*, *57*(2), 290–296. https://doi.org/10.1016/j.jpainsymman.2018.11.019

Sandilos, L. E., Lewis, K., Komaroff, E., Hammer, C. S., Scarpino, S. E., Lopez, L., Rodriguez, B., & Goldstein, B. (2015). Analysis of bilingual children's performance on the English and Spanish versions of the Woodcock-Muñoz Language Survey-R (WMLS-R). *Language Assessment Quarterly*, *12*(4), 386–408. https://doi.org/10.1080/15434303.2015.1100198

Scheurich, J. J., & Young, M. D. (1997). Coloring Epistemologies: Are Our Research Epistemologies Racially Biased? *Educational Researcher*, *26*(4), 4–16. https://doi.org/10.2307/1176879

Setodji, C. M., Peipert, J. D., & Hays, R. D. (2019). Differential item functioning of the CAHPS® In-Center Hemodialysis Survey. *Quality of Life Research: An International*

Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation, *28*(11), 3117–3135. https://doi.org/10.1007/s11136-019-02250-5

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*(2), 159–194. https://doi.org/10.1007/BF02294572

Shultz, K., Whitney, D., & Zickar, M., J. (2014). *Measurement Theory in Action: Case Studies and Exercises* (Vol. 2). https://doi.org/10.4135/9781452224749

Silverberg, J. I., Lai, J.-S., Vakharia, P. P., Patel, K., Singam, V., Chopra, R., Sacotte, R., Patel, N., Rastogi, S., Kantor, R., Hsu, D. Y., & Cella, D. (2020). Measurement properties of the Patient-Reported Outcomes Measurement Information System Itch Questionnaire item banks in adults with atopic dermatitis. *Journal of the American Academy of Dermatology*, *82*(5), 1174–1180. https://doi.org/10.1016/j.jaad.2019.11.057

Smedley, A., & Smedley, B., D. (2005). Race as biology is fiction, racism as a social problem is real: Anthropolog... *AMerican Psychologist*, *60*(1), 16–26.

Snijder, M. B., Galenkamp, H., Prins, M., Derks, E. M., Peters, R. J. G., Zwinderman, A. H., & Stronks, K. (2017). Cohort profile: The Healthy Life in an Urban Setting (HELIUS) study in Amsterdam, The Netherlands. *BMJ Open*, *7*(12), e017873. https://doi.org/10.1136/bmjopen-2017-017873

Stevanovic, D., Bagheri, Z., Atilola, O., Vostanis, P., Stupar, D., Moreira, P., Franic, T., Davidovic, N., Knez, R., Nikšić, A., Dodig-Ćurković, K., Avicenna, M., Multazam Noor, I., Nussbaum, L., Deljkovic, A., Aziz Thabet, A., Petrov, P., Ubalde, D., Monteiro, L. A., & Ribas, R. (2017). Cross-cultural measurement invariance of the Revised Child Anxiety

and Depression Scale across 11 world-wide societies. *Epidemiology and Psychiatric Sciences, 26*(4), 430–440. https://doi.org/10.1017/S204579601600038X

Stone, M. D., Matheson, B. E., Leventhal, A. M., & Boutelle, K. N. (2020). Development and validation of a short form Children's Power of Food Scale. *Appetite, 147*, 104549. https://doi.org/10.1016/j.appet.2019.104549

Suyemoto, K. L., Trimble, J. E., Cokley, K. O., Neville, H. A., Mattar, S., & Speight, S. L. (2019). *Race and Ethnicity Guidelines in Psychology: Promoting Responsiveness and Equity*. American Psychological Association. http://www.apa.org/about/policy/race-and-ethnicity-in-psychology.pdf

Tate, W. F. (1994). From Inner City to Ivory Tower: Does My Voice Matter in the Academy? *Urban Education*, *29*(3), 245–269. https://doi.org/10.1177/0042085994029003002

Teresi J. A., Fleishman J. A. (2007). Differential item functioning and health assessment. Quality of Life Research, 16(Suppl. 1), 33-42.

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Ramirez, M., & Kim, G. (2016a). Measurement equivalence of the Patient Reported Outcomes Measurement Information System® (PROMIS®) Anxiety short forms in ethnically diverse groups. *Psychological Test and Assessment Modeling*, *58*(1), 183–219.

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Ramirez, M., & Kim, G. (2016b). Psychometric properties and performance of the Patient Reported Outcomes Measurement Information System® (PROMIS®) Depression short forms in ethnically diverse groups. *Psychological Test and Assessment Modeling*, *58*(1), 141–181.

Teresi, J. A., Ocepek-Welikson, K., Ramirez, M., Fieo, R., Fulmer, T., & Gurland, B. J. (2018). Development of a Short-Form of the Medication Management Test: Evaluation of

Dimensionality, Reliability, Information and Measurement Equivalence Using Latent Variable Models. *Journal of Nursing Measurement*, *26*(3), 483–511. https://doi.org/10.1891/1061-3749.26.3.483

Teresi, J. A., Ocepek-Welikson, K., Ramirez, M., Kleinman, M., Ornstein, K., & Siu, A. (2015). Evaluation of measurement equivalence of the Family Satisfaction with the End-of-Life Care in an ethnically diverse cohort: Tests of differential item functioning. *Palliative Medicine*, *29*(1), 83–96. https://doi.org/10.1177/0269216314545802

Teresi, J. A., Ocepek-Welikson, K., Toner, J. A., Kleinman, M., Ramirez, M., Eimicke, J. P., Gurland, B. J., & Siu, A. (2017). Methodological issues in measuring subjective Well-Being and Quality-of-Life: Applications to assessment of affect in older, chronically and cognitively impaired, ethnically diverse groups using the Feeling Tone Questionnaire. *Applied Research in Quality of Life*, *12*(2), 251–288. https://doi.org/10.1007/s11482-017-9516-9

Terwee, C. B., Crins, M. H. P., Boers, M., de Vet, H. C. W., & Roorda, L. D. (2019). Validation of two PROMIS item banks for measuring social participation in the Dutch general population. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation*, *28*(1), 211–220. https://doi.org/10.1007/s11136-018-1995-0

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In *Differential item functioning.* (pp. 67–113). Lawrence Erlbaum Associates, Inc. https://doi.org/10.1075/z.62.13kok

Torres, J. B., & Colón, G. A. T. (2015). Racial Experience as an Alternative Operationalization of Race. *Human Biology*, *87*(4), 306–312.

Tucker, C. A., Bevans, K. B., Becker, B. D., Teneralli, R., & Forrest, C. B. (2020). Development of the PROMIS Pediatric Physical Activity Item Banks. *Physical Therapy*, *100*(8), 1393–1410. https://doi.org/10.1093/ptj/pzaa074

Valencia, R. R. (Ed.). (1997). *The evolution of deficit thinking: Educational thought and practice*. Falmer Press.

van Zyl, M. A., Studts, C., & Pahl, K. (2015). Precision across race, age and gender of a HIV Risk Screen for adolescents and young adults. *Social Work in Public Health*, *30*(3), 260–271.

Ward, J. T., Link, N. W., & Taylor, R. B. (2017). New windows into a broken construct: A multilevel factor analysis and DIF assessment of perceived incivilities. *Journal of Criminal Justice*, *51*, 74–88. https://doi.org/10.1016/j.jcrimjus.2017.06.004

Ward, J. T., Ray, J. V., & Fox, K. A. (2018). Exploring differences in self-control across sex, race, age, education, and language: Considering a bifactor MIMIC model. *Journal of Criminal Justice*, *56*, 29–42. https://doi.org/10.1016/j.jcrimjus.2017.09.006

Wiesner, M., Windle, M., Kanouse, D. E., Elliott, M. N., & Schuster, M. A. (2015). DISC Predictive Scales (DPS): Factor structure and uniform differential item functioning across gender and three racial/ethnic groups for ADHD, conduct disorder, and oppositional defiant disorder symptoms. *Psychological Assessment*, *27*(4), 1324–1336. https://doi.org/10.1037/pas0000101

Wijsen, L. D., & Borsboom, D. (2021). Perspectives on Psychometrics Interviews with 20 Past Psychometric Society Presidents. *Psychometrika*, *86*(1), 327–343. https://doi.org/10.1007/s11336-021-09752-7

Xu, R. H., Wong, E. L., Lu, S. Y., Zhou, L., Chang, J., & Wang, D. (2020). Validation of the Toronto Empathy Questionnaire (TEQ) among medical students in China: Analyses using three psychometric methods. *Frontiers in Psychology, 11*, 810. https://doi.org/10.3389/fpsyg.2020.00810

Zuberi, T. (2001). *Thicker than blood: How racial statistics lie*. University of Minnesota Press.

Zuckerman, M. (1990). Some dubious premises in research and theory on racial differences: Scientific, social, and ethical issues. *American Psychologist, 45*(12), 1297–1303.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF)*. Directorate of Human Resources Research and Evaluation National Defense Headquarters.

Zumbo, B. D. (2007). Three Generations of DIF Analyses: Considering Where It Has Been, Where It Is Now, and Where It Is Going. *Language Assessment Quarterly*, *4*(2), 223–233. https://doi.org/10.1080/15434300701375832

Zwick, R. (2012). A Review of Ets Differential Item Functioning Assessment Procedures: Flagging Rules, Minimum Sample Size Requirements, and Criterion Refinement. *ETS Research Report Series*, *2012*(1), i–30. https://doi.org/10.1002/j.2333-8504.2012.tb02290.

APPENDIX A: INTERVIEW PROTOCOL

Hello, my name is ‘Malitšitso Moteane, and I am a PhD student at the University of North Carolina at Greensboro. Thank you so much for taking the time to meet with me to reflect on your study _____ (name of study) that included race and/or ethnicity as a grouping variable in DIF analyses.

This interview will be kept confidential, and your responses will only be used for research purposes. Any reporting of these interviews will be done in an aggregate form, and your name and study title will not be made public in the study reporting of findings. The information sheet for this study provides additional information regarding the purpose of this study.

I am hoping to audio/video record these interviews in order to preserve the richness of our conversation for analysis, but all recordings will be deleted once transcripts are generated and cleaned. Do you mind if I begin recording?

Rationale

1. Why did you/your research team choose to conduct DIF analyses?
2. Why did you choose race and/or ethnicity as a grouping variable

Procedures

3. How did you/your team select the matching variable for the DIF analysis?
4. How did you/your research team conceptualize the race and/or ethnicity variable?
5. How did you assign participants to each racial/ethnic group?

Interpretation

6. Can you please talk me through your interpretation of the DIF analysis findings. (PROBE: How did your team interpret the direction of DIF?)

Use of DIF findings

7. Now that the article has been published, it has been cited by other scholars, (e.g. _ _). Here are some quotes from articles that cited your work. … What are your thoughts on what I am sharing with you? Do you agree with this characterization?

8. What do you think about how your work is being used?

9. Looking back at how other people have been using your findings, what would you have done differently?

10. Knowing the impact of your research and how other people use these types of research findings, what ideas do you have for future studies of DIF that use race and/or ethnicity as a grouping variable?

APPENDIX B: RECRUITMENT MATERIALS

Dear [Title and Name of 1st Author],

I hope this email finds you well. I am writing to you as I recently came across and had the privilege of working with your article, "[Insert Article Title]," published in the esteemed [Insert Peer-Reviewed Journal Name] in [Insert Publication Year] in Phase 1 of the research study 'Critically exploring the use of race and ethnicity as grouping variable in studies that use or include differential item functioning analyses.'. Your work stood out to me for its significant impact in [Insert Way in Which It Was Important] and has been cited an impressive [Insert Number of Times] times since its publication.

As this research study progresses into Phase 2, we are particularly interested in delving into the operationalization of race and/or ethnicity when used as grouping variables in Differential Item Functioning (DIF) analyses. Your paper, with its valuable insights, aligns closely with our research objectives.

At this point, I would like to invite you to participate in a process that considers the significance of the DIF findings in your paper as they are used in forward citations, especially if they align with the intentions of your original article. I am excited to extend to you the opportunity to participate in one of two ways:

1: A **once-off reflective interview via Zoom** in the next few weeks. The interview should last approximately 1 hour and will involve sharing your rationale for the study, conceptualization of

the race and/or ethnicity variable(s) as well as a reflection on how your study findings have been used in forward citations. Please click here (hyperlink to Calendy) to schedule a time for your interview.

2: **Asynchronous reflection via email** where you will receive a set of 7 reflection prompts related to your rationale for the study, conceptualization of the race and/or ethnicity variable(s) as well as a reflection on how your study findings have been used in forward citations. Through a series of structured email communications, you'll have the opportunity to share your insights and reflections at your own pace.

For both options, your participation is entirely voluntary, and we will maintain the confidentiality of your responses. Your expertise and insights would be invaluable in advancing our research. I am eager to discuss how your work has influenced the field and how it is utilized in subsequent research. Your participation would greatly contribute to the success of our project.

I look forward to the possibility of your involvement, and I'm available to answer any questions you may have. Please feel free to contact me on this email address.

Warm regards,

Dear _____ (1st authors name and title),

My name is ‘Malitšitso Moteane. I am a PhD candidate at the University of North Carolina at Greensboro, and I am conducting my dissertation study titled “Critically exploring the use of race and ethnicity as grouping variable in studies that use or include differential item functioning analyses.”

I would like to invite you (or any member of your research team) to a reflective interview via Zoom in the next few weeks. The interview should last approximately 1 hour. In preparation for our interview, I have performed a forward citation search of your DIF study and have attached a full list of forward citations and a selection of excerpts for reflection in the interview. My hope with this study and specifically your reflections on the work that cites your research is to generate new knowledge about how findings from DIF studies/studies that include DIF analysis are used in forward citations.

The purpose of the study is to explore trends in the conceptualization of race and ethnicity as grouping variables, the theoretical framing, interpretation, and onward use of findings in differential item functioning (DIF) studies that use race/ethnicity as a grouping variable. This is an explanatory sequential mixed methods study. The first phase of this study was a systematic literature review of published peer reviewed studies that included differential item functioning (DIF) analyses where race and/or ethnicity were used as grouping variable. Your study has been selected based on the findings from the systematic literature review. In particular, the variable ‘interpretation of DIF findings’ was used to categorize DIF studies that used race and/or ethnicity as grouping variables into three categories. Your research study was categorized as (no interpretation/interpretation of DIF findings but not linked to race/ethnicity grouping variable/ full interpretation of DIF findings in relations to race/ethnicity grouping

variable). THREE out of the xx studies in this category were selected for interviews.

In the interview, you will be invited to share your rationale for the study, conceptualization of the race and/or ethnicity variable as well as a reflection on how your study findings have been used in forward citations.

If you are interested, please click here (hyperlink to Calendy) to schedule a time for your interview. I look forward to hearing from you.

APPENDIX C: CONSENT FORM

UNIVERSITY OF NORTH CAROLINA AT GREENSBORO

CONSENT TO ACT AS A HUMAN PARTICIPANT

Project Title: Critically exploring the use of race and ethnicity as grouping variables in studies that use or include differential item functioning analyses

Principal Investigator and Faculty Advisor (if applicable): ‘Malitšitso Moteane (PhD Candidate) Dr. Micheline Chalhoub-Deville (Faculty Advisor)

Participant's Name:

What are some general things you should know about research studies?

You are being asked to take part in a research study. Your participation in the study is voluntary. You may choose not to join, or you may withdraw your consent to be in the study, for any reason, without penalty.

Research studies are designed to obtain new knowledge. This new information may help people in the future. There may not be any direct benefit to you for being in the research study. No risks are readily apparent from your participation in this research study. If you choose not to be in the study or leave the study before it is done, it will not affect your relationship with the researcher or the University of North Carolina at Greensboro.

Details about this study are discussed in this consent form. It is important that you understand this information so that you can make an informed choice about being in this research study.

You will be given a copy of this consent form. If you have any questions about this study at any time, you should ask the researchers named in this consent form. Their contact information is below.

What is the study about?

This is a research project. Your participation is voluntary. The purpose of this study is to explore trends in the conceptualization of race and ethnicity as grouping variables, the theoretical framing, interpretation and onward use of findings in differential item functioning (DIF) studies that use race/ethnicity as a grouping variable.

Why are you asking me?

This is an explanatory sequential mixed methods study. The first phase of this study was a systematic literature review of published peer reviewed studies that included differential item functioning (DIF) analyses where race and/or ethnicity were used as grouping variable. You have been selected based on the findings from the systematic literature review. In particular, the variable 'interpretation of DIF findings' was used to categorize DIF studies that used race and/or ethnicity as grouping variables. Your research study was categorized as (no interpretation/interpretation of DIF findings but not linked to race/ethnicity grouping variable/ full interpretation of DIF findings in relations to race/ethnicity grouping variable). Three out of the xx studies in the (no interpretation/interpretation of DIF findings but not linked to

race/ethnicity grouping variable/ full interpretation of DIF findings in relations to race/ethnicity grouping variable) were selected for interviews.

What will you ask me to do if I agree to be in the study?

Your participation in this study will involve sitting for ONE reflective interview that should last no longer than one hour.

Is there any audio/video recording?

Interviews will be audio / video recorded with your permission for the purpose of producing an accurate transcript for data analysis. Because your voice will be potentially identifiable by anyone who hears the recording, your confidentiality for things you say on the recording cannot be guaranteed although the researcher will try to limit access to the recording as described below.

What are the risks to me?

The Institutional Review Board at the University of North Carolina at Greensboro has determined that participation in this study poses no identifiable risk to participants.

If you have questions, want more information, or have suggestions, please contact Malitšitso Moteane (m_motean@uncg.edu) who may be reached at (336) 604-9675 OR her Faculty Advisor Dr. Micheline Chalhoub-Deville (mbchalho@uncg.edu or chalhoub-deville@uncg.edu).

If you have any concerns about your rights, how you are being treated, concerns or complaints about this project or benefits or risks associated with being in this study please contact the Office of Research Integrity at UNCG toll-free at (855)-251-2351.

Are there any benefits to society as a result of me taking part in this research?

The findings of this study may inform researchers who use race and/or ethnicity in their DIF studies on current conceptualizations of the variables and considerations that they need to make when using race and ethnicity as grouping variables.

Are there any benefits to *me* for taking part in this research study?

As part of the interview protocol for this study a forward citation search will be conducted for the DIF research study you authored and shared with you. This will include a count of forward citation as at _____ (date one week before the interview) and excerpts from said forward citations.

Will I get paid for being in the study? Will it cost me anything?

There are no costs to you or payments made for participating in this study.

How will you keep my information confidential?

Video/audio recordings will be stored in a secure, university approved cloud storage password protected folder until transcripts are produced and cleaned after which they will be permanently deleted. Transcripts will be de-identified and stored in the same university approved password protected cloud storage folder that will only be accessed by the researcher and dissertation

advisory committee. Your data will be destroyed one calendar year after completion of the study. All information obtained in this study is strictly confidential unless disclosure is required by law.

Absolute confidentiality of data provided through the Internet cannot be guaranteed due to the limited protections of Internet access. Please be sure to close your browser when finished so no one will be able to see what you have been doing.

Will my de-identified data be used in future studies?

Your data will be destroyed one calendar year after completion of the study. De-identified data will not be stored long term and will not be used in future research projects.

What if I want to leave the study?

You have the right to refuse to participate or to withdraw at any time, without penalty. If you do withdraw, it will not affect you in any way. If you choose to withdraw, you may request that any of your data which has been collected be destroyed unless it is in a de-identifiable state. The investigators also have the right to stop your participation at any time. This could be because of reasons such as an unexpected reaction, or you have failed to follow instructions, or because the entire study has been stopped.

What about new information/changes in the study?

If significant new information relating to the study becomes available which may relate to your willingness to continue to participate, this information will be provided to you.

Voluntary Consent by Participant:

By signing this consent form/completing this survey/activity (used for an IRB-approved waiver of signature) you are agreeing that you read, or it has been read to you, and you fully understand the contents of this document and are openly willing consent to take part in this study. All of your questions concerning this study have been answered. By signing this form, you are agreeing that you are 18 years of age or older and are agreeing to participate, in this study described to you by _____.

Signature: _____ Date: _____

APPENDIX D: LIST OF SAMPLED ARTICLES

Abdin, E., Sagayadevan, V., Vaingankar, J. A., Picco, L., Chong, S. A., & Subramaniam, M. (2018). A non-parametric item response theory evaluation of the CAGE instrument among older adults. *Substance Use & Misuse*, *53*(3), 391–399. https://doi.org/10.1080/10826084.2017.1332645

Abdin, E., Subramaniam, M., Chan, A., Chen, J.-A., Chong, C. L., Wang, C., Lee, M., & Gan, S. L. (2019). iWorkHealth: An instrument to identify workplace psychosocial risk factors for a multi-ethnic Asian working population. *PLoS ONE*, *14*(8). https://doi.org/10.1371/journal.pone.0220566

Abdin, E., Subramaniam, M., Picco, L., Pang, S., Vaingankar, J. A., Shahwan, S., Sagayadevan, V., Zhang, Y., & Chong, S. A. (2017). The importance of considering differential item functioning in investigating the impact of chronic conditions on health-related quality of life in a multi-ethnic Asian population. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation*, *26*(4), 823–834. https://doi.org/10.1007/s11136-016-1418-z

Abrajano, M. (2015). Reexamining the "racial gap" in political knowledge. *The Journal of Politics*, *77*(1), 44–54. https://doi.org/10.1086/678767

Adams, D., Joo, M. T. H., Sumintono, B., & Pei, O. S. (2020). Blended learning engagement in higher education institutions: A differential item functioning analysis of students' backgrounds. *Malaysian Journal of Learning and Instruction*, *17*(Number 1), 133–158. https://doi.org/10.32890/mjli2020.17.1.6

Adams, D., Sumintono, Bambang, Mohamed, Ahmed, & Noor, Nur Syafika Mohamad. (2018). E-Learning Readiness among Students of Diverse Backgrounds in a Leading Malaysian Higher Education Institution. *Malaysian Journal of Learning and Instruction*, *15*(No 2), 227–256.

Aksu-Dunya, B. (2018). Psychometric Properties and Differential Item Functioning of a Web-Based Assessment of Children's Social Perspective-Taking. *Journal of Applied Measurement, 19*(1), 93–105.

Aksu-Dunya, B., McKown, C., & Smith, E. (2020). Psychometric Properties and Differential Item Functioning of a Web-Based Assessment of Children's Emotion Recognition Skill. *Journal of Psychoeducational Assessment*, *38*(5), 627–641. https://doi.org/10.1177/0734282919881919

Armenta, B. E., & Cooper, M. L. (2019). The Rutgers Alcohol Problem Index: Measurement equivalence among college students in the US and Mexico. *Psychological Assessment*, *31*(1), 1–14. https://doi.org/10.1037/pas0000608

Barau, A. S. (2015). Perceptions and contributions of households towards sustainable urban green infrastructure in Malaysia. *Habitat International*, *47*, 285–297. https://doi.org/10.1016/j.habitatint.2015.02.003

Bowe, A. (2017). The cultural fairness of the 12-item General Health Questionnaire among diverse adolescents. *Psychological Assessment, 29*(1), 87–97. https://doi.org/10.1037/pas0000323

Bowe, A. (2019). Moving Toward More Conclusive Measures of Sociocultural Adaptation for Ethnically Diverse Adolescents in England. *Canadian Journal of School Psychology*, *34*(1), 56–72. https://doi.org/10.1177/0829573517739392

Cartwright, J. K., Desmarais, S. L., Grimm, K. J., Meade, A. W., & Van Dorn, R. A. (2020). Psychometric Properties of the MacArthur Community Violence Screening Instrument. *International Journal of Forensic Mental Health*, *19*(3), 253–268. https://doi.org/10.1080/14999013.2020.1718246

Chakawa, A., Butler, R. C., & Shapiro, S. K. (2015). Examining the psychometric validity of the Multigroup Ethnic Identity Measure-Revised (MEIM-R) in a community sample of African

American and European American adults. *Cultural Diversity and Ethnic Minority Psychology*, *21*(4), 643–648. https://doi.org/10.1037/cdp0000025

Charles, P., Belisle, M., Tonita, K., & Smith, J. (2015). Help Me Tell My Story: Development of an Oral Language Measurement Scale. *Journal of Applied Measurement*, *16*(3), 278–297.

Chen, C.-Y., Squires, J., Chen, C.-I., Wu, R., & Xie, H. (2020). The Adaptation and Psychometric Examination of a Social-Emotional Developmental Screening Tool in Taiwan. *Early Education and Development*, *31*(1), 27–46.

Chen, H. F., & Zhu, J. (2017). Optimal items for assessing parental involvement across different groups during middle childhood. *Journal of Child and Family Studies*, *26*(11), 2999–3012. https://doi.org/10.1007/s10826-017-0809-2

Chen, T.-A., O'Connor, T. M., Hughes, S. O., Beltran, A., Baranowski, J., Diep, C., & Baranowski, T. (2015). Vegetable parenting practices scale. Item response modeling analyses. *Appetite*, *91*, 190–199. https://doi.org/10.1016/j.appet.2015.04.048

Cicero, D. C., Martin, E. A., & Krieg, A. (2019). Differential Item Functioning of the Full and Brief Wisconsin Schizotypy Scales in Asian, White, Hispanic, and Multiethnic Samples and Between Sexes. *Assessment*, *26*(6), 1001–1013. https://doi.org/10.1177/1073191117719509

Cordier, R., Munro, N., Wilkes-Gillan, S., Speyer, R., Parsons, L., & Joosten, A. (2019). Applying Item Response Theory (IRT) Modeling to an Observational Measure of Childhood Pragmatics: The Pragmatics Observational Measure-2. *Frontiers in Psychology*, *10*, 408. https://doi.org/10.3389/fpsyg.2019.00408

Crowder, M. K., Gordon, R. A., Brown, R. D., Davidson, L. A., & Domitrovich, C. E. (2019). Linking Social and Emotional Learning Standards to the WCSD Social-Emotional Competency Assessment: A Rasch Approach. *School Psychology*, *34*(3), 281–295.

Curran, E., Adamson, G., Rosato, M., Cock, P., & Leavey, G. (2018). Profiles of childhood trauma and psychopathology: US National Epidemiologic Survey. *Social Psychiatry and Psychiatric Epidemiology: The International Journal for Research in Social and Genetic Epidemiology and Mental Health Services*, *53*(11), 1207–1219. https://doi.org/10.1007/s00127-018-1525-y

Devine, R. T., & Hughes, C. (2016). Measuring theory of mind across middle childhood: Reliability and validity of the Silent Films and Strange Stories tasks. *Journal of Experimental Child Psychology*, *149*, 23–40. https://doi.org/10.1016/j.jecp.2015.07.011

Dmitrieva, N. O., Fyffe, D., Mukherjee, S., Fieo, R., Zahodne, L. B., Hamilton, J., Potter, G. G., Manly, J. J., Romero, H. R., Mungas, D., & Gibbons, L. E. (2015). Demographic characteristics do not decrease the utility of depressive symptoms assessments: Examining the practical impact of item bias in four heterogeneous samples of older adults: Differential item function in depressive symptoms. *International Journal of Geriatric Psychiatry*, *30*(1), 88–96. https://doi.org/10.1002/gps.4121

Du Plessis, G. A., & De Bruin, G. P. (2015). Using Rasch modelling to examine the international personality item pool (IPIP) values in action (VIA) measure of character strengths. *Journal of Psychology in Africa*, *25*(6), 512–521. https://doi.org/10.1080/14330237.2015.1124603

DuPaul, G. J., Fu, Q., Anastopoulos, A. D., Reid, R., & Power, T. J. (2020). ADHD parent and teacher symptom ratings: Differential item functioning across gender, age, race, and ethnicity. *Journal of Abnormal Child Psychology*, *48*(5), 679–691. https://doi.org/10.1007/s10802-020-00618-7

Ehrich, J., Howard, S. J., Mu, C., & Bokosmaty, S. (2016). A comparison of Chinese and Australian university students' attitudes towards plagiarism. *Studies in Higher Education*, *41*(2), 231–246. https://doi.org/10.1080/03075079.2014.927850

Farrington, A. L., & Lonigan, C. J. (2015). Examining the Measurement Precision and Invariance of the Revised Get Ready to Read! *Journal of Learning Disabilities*, *48*(3), 227–238. https://doi.org/10.1177/0022219413495568

Fieo, R., Ocepek-Welikson, K., Kleinman, M., Eimicke, P., Crane, P. K., Cella, D., & Teresi, J. A. (2016). Measurement equivalence of the Patient Reported Outcomes Measurement Information System® (PROMIS®) Applied Cognition – General Concerns, short forms in ethnically diverse groups. *Psychological Test and Assessment Modeling*, *58*(2), 255–307.

Filshtein, T., Chan, M., Mungas, D., Whitmer, R., Fletcher, E., DeCarli, C., & Farias, S. (2020). Differential Item Functioning of the Everyday Cognition (ECog) Scales in Relation to Racial/Ethnic Groups. *Journal of the International Neuropsychological Society*, *26*(5), 515–526. https://doi.org/10.1017/S1355617719001437

Flanagan, A. M., Cormier, D. C., & Bulut, O. (2020). Achievement may be rooted in teacher expectations: Examining the differential influences of ethnicity, years of teaching, and classroom behaviour. *Social Psychology of Education: An International Journal*, *23*(6), 1429–1448. https://doi.org/10.1007/s11218-020-09590-y

Forrest, C. B., Devine, J., Bevans, K. B., Becker, B. D., Carle, A. C., Teneralli, R. E., Moon, J., Tucker, C. A., & Ravens-Sieberer, U. (2018). Development and psychometric evaluation of the PROMIS Pediatric Life Satisfaction item banks, child-report, and parent-proxy editions. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation*, *27*(1), 217–234. https://doi.org/10.1007/s11136-017-1681-7

Forrest, C. B., Ravens-Sieberer, U., Devine, J., Becker, B. D., Teneralli, R. E., Moon, J., Carle, A. C., Tucker, C. A., & Bevans, K. B. (2018). Development and evaluation of the PROMIS® Pediatric Positive Affect item bank, child-report and parent-proxy editions. *Journal of Happiness Studies:*

An Interdisciplinary Forum on Subjective Well-Being, *19*(3), 699–718. https://doi.org/10.1007/s10902-016-9843-9

French, B. F., & Vo, T. T. (2020). Differential Item Functioning of a Truancy Assessment. *Journal of Psychoeducational Assessment*, *38*(5), 642–648. https://doi.org/10.1177/0734282919863215

Galenkamp, H., Stronks, K., Mokkink, L. B., & Derks, E. M. (2018). Measurement invariance of the SF-12 among different demographic groups: The HELIUS study. *PLoS ONE*, *13*(9). https://doi.org/10.1371/journal.pone.0203483

Galenkamp, H., Stronks, K., Snijder, M. B., & Derks, E. M. (2017). Measurement invariance testing of the PHQ-9 in a multi-ethnic population in Europe: The HELIUS study. *BMC Psychiatry*, *17*. https://doi.org/10.1186/s12888-017-1506-9

Gay, C. L., Kottorp, A., Lerdal, A., & Lee, K. A. (2016). Psychometric limitations of the Center for Epidemiologic Studies-Depression scale for assessing depressive symptoms among adults with HIV/AIDS: A Rasch analysis. *Depression Research and Treatment*, *2016*. https://doi.org/10.1155/2016/2824595

Geldenhuys, M., & Bosch, A. (2020). A Rasch adapted version of the 30-item Bem Sex Role Inventory (BSRI). *Journal of Personality Assessment*, *102*(3), 428–439. https://doi.org/10.1080/00223891.2018.1527343

Goetz, C. G., Liu, Y., Stebbins, G. T., Wang, L., Tilley, B. C., Teresi, J. A., Merkitch, D., & Luo, S. (2016). Gender-, age-, and race/ethnicity-based differential item functioning analysis of the movement disorder society-sponsored revision of the Unified Parkinson's disease rating scale: DIF Analysis Of MDS-UPDRS. *Movement Disorders*, *31*(12), 1865–1873. https://doi.org/10.1002/mds.26847

Gross, G. M., Kwapil, T. R., Raulin, M. L., Silvia, P. J., & Barrantes-Vidal, N. (2018). The multidimensional schizotypy scale-brief: Scale development and psychometric properties. *Psychiatry Research, 261*, 7–13. https://doi.org/10.1016/j.psychres.2017.12.033

Hahn, E. A., Kallen, M. A., Jensen, R. E., Potosky, A. L., Moinpour, C. M., Ramirez, M., Cella, D., & Teresi, J. A. (2016). Measuring social function in diverse cancer populations: Evaluation of measurement equivalence of the Patient Reported Outcomes Measurement Information System® (PROMIS®) Ability to Participate in Social Roles and Activities short form. *Psychological Test and Assessment Modeling, 58*(2), 403–421.

Haroz, E. E., Bolton, P., Gross, A., Chan, K. S., Michalopoulos, L., & Bass, J. (2016). Depression symptoms across cultures: An IRT analysis of standard depression symptoms using data from eight countries. *Social Psychiatry and Psychiatric Epidemiology: The International Journal for Research in Social and Genetic Epidemiology and Mental Health Services*, *51*(7), 981–991. https://doi.org/10.1007/s00127-016-1218-3

Harpole, J. K., Levinson, C. A., Woods, C. M., Rodebaugh, T. L., Weeks, J. W., Brown, P. J., Heimberg, R. G., Menatti, A. R., Blanco, C., Schneier, F., & Liebowitz, M. (2015). Assessing the straightforwardly-worded Brief Fear of Negative Evaluation Scale for differential item functioning across gender and ethnicity. *Journal of Psychopathology and Behavioral Assessment, 2,* 306–317. https://doi.org/10.1007/s10862-014-9455-9

Harris, K. M., & Aboujaoude, E. (2016). Online Friendship, Romance, and Sex: Properties and Associations of the Online Relationship Initiation Scale. *Cyberpsychology, Behavior, and Social Networking, 19*(8), 487–493. https://doi.org/10.1089/cyber.2016.0164

Harris, K. M., Syu, J.-J., Lello, O. D., Chew, Y. L. E., Willcox, C. H., & Ho, R. H. M. (2015). The ABC's of Suicide Risk Assessment: Applying a Tripartite Approach to Individual Evaluations. *PLOS ONE*, *10*(6), e0127442. https://doi.org/10.1371/journal.pone.0127442

Hasnain, M., Gruss, V., Keehn, M., Peterson, E., Valenta, A. L., & Kottorp, A. (2017). Development and validation of a tool to assess self-efficacy for competence in interprofessional collaborative practice. *Journal of Interprofessional Care*, *31*(2), 255–262. https://doi.org/10.1080/13561820.2016.1249789

Hawes, S. W., Byrd, A. L., Kelley, S. E., Gonzalez, R., Edens, J. F., & Pardini, D. A. (2018). Psychopathic features across development: Assessing longitudinal invariance among Caucasian and African American youths. *Journal of Research in Personality*, *73*, 180–188. https://doi.org/10.1016/j.jrp.2018.02.003

Heafner, T. L., & Fitchett, P. G. (2018). US history content knowledge and associated effects of race, gender, wealth, and urbanity: Item Response Theory (IRT) modeling of NAEP-USH achievement. *Journal of Social Studies Research*, *42*(1), 11–25. https://doi.org/10.1016/j.jssr.2017.01.001

Hong, I., Velozo, C. A., Li, C.-Y., Romero, S., Gruber-Baldini, A. L., & Shulman, L. M. (2016). Assessment of the psychometrics of a PROMIS item bank: Self-efficacy for managing daily activities. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation*, *25*(9), 2221–2232. https://doi.org/10.1007/s11136-016-1270-1

Hu, J., Serovich, J., Brown, M., Kimberly, J., & Chen, Y.-H. (2017). Psychometric Evaluation of the HIV Disclosure Belief Scale: A Rasch Model Approach. *AIDS & Behavior*, *21*(1), 174–183. https://doi.org/10.1007/s10461-016-1478-7

Jang, Y., Powers, D. A., Yoon, H., Rhee, M.-K., Park, N. S., & Chiriboga, D. A. (2018). Measurement equivalence of English versus native language versions of the Kessler 6 (K6) Scale: An examination in three Asian American groups. *Asian American Journal of Psychology*, *9*(3), 211–216. https://doi.org/10.1037/aap0000110

Janulis, P., Newcomb, M. E., Sullivan, P., & Mustanski, B. (2018). Evaluating HIV knowledge questionnaires among men who have sex with men: A multi-study item response theory analysis. *Archives of Sexual Behavior*, *47*(1), 107–119. https://doi.org/10.1007/s10508-016-0910-4

Jean-Pierre, P., Cheng, Y., & Paxton, R. J. (2020). Item-level psychometrics of a brief self-reported memory problem screening measure in breast cancer survivors. *Acta Oncologica*, *59*(3), 358–364. https://doi.org/10.1080/0284186X.2019.1687935

Jensen, R. E., King-Kallimanis, B. L., Sexton, E., Reeve, B. B., Moinpour, C. M., Potosky, A. L., Lobo, T., & Teresi, J. A. (2016). Measurement properties of PROMIS Sleep Disturbance short forms in a large, ethnically diverse cancer cohort. *Psychological Test and Assessment Modeling*, *58*(2).

Jones, R. N., Tommet, D., Ramirez, M., Jensen, R., & Teresi, J. A. (2016). Differential item functioning in Patient Reported Outcomes Measurement Information System® (PROMIS®) Physical Functioning short forms: Analyses across ethnically diverse groups. *Psychological Test and Assessment Modeling*, *58*(2), 371–402.

Kim, G., DeCoster, J., Bryant, A. N., & Ford, K. L. (2016). Measurement Equivalence of the K6 Scale. *Assessment*, *23*(6), 758–768. https://doi.org/10.1177/1073191115599639

Kwan, Y. H., Uy, E. J., Bautista, D. C., Xin, X., Xiao, Y., Lee, G. L., Subramaniam, M., Vaingankar, J. A., Chan, M. F., Kumar, N., Cheung, Y. B., Chua, T. S. J., & Thumboo, J. (2019a). Development and calibration of a novel positive mindset item bank to measure health-related

quality of life (HRQoL) in Singapore. *PLoS ONE*, *14*(7). https://doi.org/10.1371/journal.pone.0220293

Kwan, Y. H., Uy, E. J., Bautista, D. C., Xin, X., Xiao, Y., Lee, G. L., Subramaniam, M., Vaingankar, J. A., Chan, M. F., Kumar, N., Cheung, Y. B., Chua, T. S. J., & Thumboo, J. (2019b). Development and calibration of a novel social relationship item bank to measure health-related quality of life (HRQoL) in Singapore. *Health and Quality of Life Outcomes*, *17*. https://doi.org/10.1186/s12955-019-1150-9

Lambert, M. C., Garcia, A. G., Epstein, M. H., & Cullinan, D. (2018). Differential Item Functioning of the Emotional and Behavioral Screener for Caucasian and African American Elementary School Students. *Journal of Applied School Psychology*, *34*(3), 201–214. https://doi.org/10.1080/15377903.2017.1345815

Lambert, M. C., January, S.-A. A., Cress, C. J., Epstein, M. H., & Cullinan, D. (2018). Differential Item Functioning across Race and Ethnicity for the Emotional and Behavioral Screener. *School Psychology Quarterly*, *33*(3), 399–407.

Lange, A. M. C., Scholte, R. H. J., van Geffen, W., Timman, R., Busschbach, J. J. V., & van der Rijken, R. E. A. (2016). The lack of cross-national equivalence of a Therapist Adherence Measure (TAM-R) in multisystemic therapy (MST). *European Journal of Psychological Assessment*, *32*(4), 312–325. https://doi.org/10.1027/1015-5759/a000262

Leung, Y. Y., Uy, E. J. B., Bautista, D. C., Pua, Y. H., Kwan, Y. H., Cheung, Y. B., Xiao, Y., Chua, T. S. J., & Thumboo, J. (2020). Calibration of a physical functioning item bank for measurement of health-related quality of life in Singapore. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation*, *29*(10), 2823–2833. https://doi.org/10.1007/s11136-020-02535-0

Li, L. Y., Meyer, M. S., Martin, E. A., Gross, G. M., Kwapil, T. R., & Cicero, D. C. (2020). Differential item functioning of the Multidimensional Schizotypy Scale and Multidimensional Scale-Brief across ethnicity. *Psychological Assessment*, *32*(4), 383–393. https://doi.org/10.1037/pas0000798

Lindhiem, O., Vaughn-Coaxum, R. A., Higa, J., Harris, J. L., Kolko, D. J., & Pilkonis, P. A. (2019). Development and validation of the Knowledge of Effective Parenting Test (KEPT) in a nationally representative sample. *Psychological Assessment*, *31*(6), 781–792. https://doi.org/10.1037/pas0000699

Loyd, A. B., Derlan, C. L., Smith, E. V., Norris, S. A., Richter, L. M., & Roeser, R. W. (2019). Evaluating the Psychometric Properties of a Measure of Ethnic Identity Among Black South African Youth. *Identity*, *19*(1), 44–61. https://doi.org/10.1080/15283488.2019.1566070

Lu, X., Yeo, K. J., Guo, F., & Zhao, Z. (2020). Factor structure and a multiple indicators multiple cause model of internet addiction test: The effect of socio-demographic and internet use variables. *Current Psychology: A Journal for Diverse Perspectives on Diverse Psychological Issues*, *39*(3), 769–781. https://doi.org/10.1007/s12144-019-00234-9

Martin, J. L., Colvin, K. F., Madson, M. B., Zamboanga, B. L., & Pazienza, R. (2020). Optimal assessment of protective behavioral strategies among college drinkers: An item response theory analysis. *Psychological Assessment*, *32*(4), 394–406. https://doi.org/10.1037/pas0000799

Mâsse, L. C., O'Connor, T. M., Lin, Y., Carbert, N. S., Hughes, S. O., Baranowski, T., & Beauchamp, M. R. (2020). The Physical Activity Parenting Practices (PAPP) item bank: A psychometrically validated tool for improving the measurement of physical activity parenting practices of parents of 5–12-year-old children. *The International Journal of Behavioral Nutrition and Physical Activity*, *17*. https://doi.org/10.1186/s12966-020-01036-0

Mâsse, L. C., O'Connor, T. M., Lin, Y., Hughes, S. O., Tugault-Lafleur, C. N., Baranowski, T., & Beauchamp, M. R. (2020). Calibration of the food parenting practice (FPP) item bank: Tools for improving the measurement of food parenting practices of parents of 5–12-year-old children. *The International Journal of Behavioral Nutrition and Physical Activity*, *17*. https://doi.org/10.1186/s12966-020-01049-9

McDonald, S. E., Ma, L., Green, K. E., Hitti, S. A., Cody, A. M., Donovan, C., Williams, J. H., & Ascione, F. R. (2018). Evaluation of the Parent-Report Inventory of Callous–Unemotional Traits in a Sample of Children Recruited from Intimate Partner Violence Services: A Multidimensional Rasch Analysis. *Journal of Clinical Psychology*, *74*(3), 418–441. https://doi.org/10.1002/jclp.22497

McFarland, J. L., Price, R. M., Wenderoth, M. P., Martinková, P., Cliff, W., Michael, J., Modell, H., & Wright, A. (2017). Development and Validation of the Homeostasis Concept Inventory. *CBE—Life Sciences Education*, *16*(2), ar35. https://doi.org/10.1187/cbe.16-10-0305

Miller, A. P., Merkle, E. C., Galenkamp, H., Stronks, K., Derks, E. M., & Gizer, I. R. (2019). Differential item functioning analysis of the CUDIT and relations with alcohol and tobacco use among men across five ethnic groups: The HELIUS study. *Psychology of Addictive Behaviors*, *33*(8), 697–709. https://doi.org/10.1037/adb0000521

Monterrosa-Castro, A., Portela-Buelvas, K., Oviedo, H. C., Herazo, E., & Campo-Arias, A. (2016). Differential Item Functioning of the Psychological Domain of the Menopause Rating Scale. *BioMed Research International*, *2016*, 1–4. https://doi.org/10.1155/2016/8790691

Nicholson, N. R., Feinn, R., Casey, E. A., & Dixon, J. (2020). Psychometric Evaluation of the Social Isolation Scale in Older Adults. *The Gerontologist*, *60*(7), e491–e501. https://doi.org/10.1093/geront/gnz083

Ning, H. K. (2018). A Rasch analysis of the Junior Metacognitive Awareness Inventory with Singapore students. *Measurement & Evaluation in Counseling & Development*, *51*(2), 84–91. https://doi.org/10.1080/07481756.2017.1358061

Owens, S., Kristjansson, A., & Hunte, H. E. R. (2015). A Differential Item Functional analysis by age of perceived interpersonal discrimination in a multi-racial/ethnic sample of adults. *Ethnicity & Disease*, *25*(4), 479. https://doi.org/10.18865/ed.25.4.479

Park, I. H., Rachmatullah, A., Park, I.-S., & Liu, Y. (2019). Assessment of the quality and generalizability of the revised UCLA loneliness scale in Chinese and Korean community-dwelling elderly populations using item response theory (IRT)-Rasch modeling and hybrid IRT-logistic regression. *Educational Gerontology*, *45*(10), 581–599. https://doi.org/10.1080/03601277.2019.1670908

Parkerson, H. A., Thibodeau, M. A., Brandt, C. P., Zvolensky, M. J., & Asmundson, G. J. G. (2015). Cultural-based biases of the GAD-7. *Journal of Anxiety Disorders*, *31*, 38–42. https://doi.org/10.1016/j.janxdis.2015.01.005

Pedersen, E. R., Huang, W., Dvorak, R. D., Prince, M. A., & Hummer, J. F. (2017). The Protective Behavioral Strategies for Marijuana Scale: Further examination using item response theory. *Psychology of Addictive Behaviors*, *31*(5), 548–559. https://doi.org/10.1037/adb0000271

Peipert, J. D., Bentler, P., Klicko, K., & Hays, R. D. (2018). Negligible impact of differential item functioning between Black and White dialysis patients on the Kidney Disease Quality of Life 36-item short form survey (KDQOLTM-36). *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation*, *27*(10), 2699–2707. https://doi.org/10.1007/s11136-018-1879-3

Quach, C. W., Langer, M. M., Chen, R. C., Thissen, D., Usinger, D. S., Emerson, M. A., & Reeve, B. B. (2016). Reliability and validity of PROMIS measures administered by telephone interview in a longitudinal localized prostate cancer study. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation*, *25*(11), 2811–2823. https://doi.org/10.1007/s11136-016-1325-3

Rachmatullah, A., Lee, J.-K., & Ha, M. (2020). Preservice science teachers' ecological value orientation: A comparative study between Indonesia and Korea. *The Journal of Environmental Education*, *51*(1), 14–28. https://doi.org/10.1080/00958964.2019.1633989

Reeve, B. B., Pinheiro, L. C., Jensen, R. E., Teresi, J. A., Potosky, A. L., McFatrich, M. K., Ramirez, M., & Chen, W.-H. (2016). Psychometric evaluation of the PROMIS® Fatigue measure in an ethnically and racially diverse population-based sample of cancer patients. *Psychological Test and Assessment Modeling*, *58*(1), 119–139.

Reich, H., Rief, W., Brähler, E., & Mewes, R. (2018). Cross-cultural validation of the German and Turkish versions of the PHQ-9: An IRT approach. *BMC Psychology*, *6*(1), 26. https://doi.org/10.1186/s40359-018-0238-z

Ribeiro Santiago, P. H., Nielsen, T., Smithers, L. G., Roberts, R., & Jamieson, L. (2020). Measuring stress in Australia: Validation of the Perceived Stress Scale (PSS-14) in a national sample. *Health and Quality of Life Outcomes*, *18*. https://doi.org/10.1186/s12955-020-01343-x

Rice, S. M., Parker, A. G., Mawren, D., Clifton, P., Harcourt, P., Lloyd, M., Kountouris, A., Smith, B., McGorry, P. D., & Purcell, R. (2020). Preliminary psychometric validation of a brief screening tool for athlete mental health among male elite athletes: The Athlete Psychological Strain Questionnaire. *International Journal of Sport and Exercise Psychology*, *18*(6), 850–865. https://doi.org/10.1080/1612197X.2019.1611900

Rodriguez, V. J., Shaffer, A., Are, F., Madden, A., Jones, D. L., & Kumar, M. (2019). Identification of differential item functioning by race and ethnicity in the Childhood Trauma Questionnaire. *Child Abuse & Neglect*, *94*, 104030. https://doi.org/10.1016/j.chiabu.2019.104030

Rose, A. J., Bayliss, E., Huang, W., Baseman, L., Butcher, E., García, R.-E., & Edelen, M. O. (2018). Evaluating the PROMIS-29 v20 for use among older adults with multiple chronic conditions. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation*, *27*(11), 2935–2944. https://doi.org/10.1007/s11136-018-1958-5

Rosselli, M., Tappen, R. M., & Newman, D. (2019). Semantic Interference Test: Evidence for culture and education fairness from an ethnically diverse sample in the USA. *Archives of Clinical Neuropsychology*, *34*(3), 337–349. https://doi.org/10.1093/arclin/acy037

Roth, D. L., Dilworth-Anderson, P., Huang, J., Gross, A. L., & Gitlin, L. N. (2015). Positive Aspects of Family Caregiving for Dementia: Differential Item Functioning by race. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *70*(6), 813–819. https://doi.org/10.1093/geronb/gbv034

Roy, C., Bakan, G., Li, Z., & Nguyen, T. H. (2016). Coping measurement: Creating short form of Coping and Adaptation Processing Scale using item response theory and patients dealing with chronic and acute health conditions. *Applied Nursing Research*, *32*, 73–79. https://doi.org/10.1016/j.apnr.2016.06.002

Ruchensky, J. R., Balsis, S., Edens, J. F., & Douglas, K. S. (2021). Suicidal ideation across race in a justice-involved sample: An item response theory approach. *Suicide and Life-Threatening Behavior*, *51*(3), 385–393. https://doi.org/10.1111/sltb.12717

Ruglass, L. M., Morgan-López, A. A., Saavedra, L. M., Hien, D. A., Fitzpatrick, S., Killeen, T. K., Back, S. E., & López-Castro, T. (2020). Measurement nonequivalence of the Clinician-

Administered PTSD Scale by race/ethnicity: Implications for quantifying posttraumatic stress disorder severity. *Psychological Assessment, 32*(11), 1015–1027. https://doi.org/10.1037/pas0000943

Sabatini, J., Bruce, K., Steinberg, J., & Weeks, J. (2015). SARA Reading Components Tests, RISE Forms: Technical Adequacy and Test Design, 2nd Edition. *ETS Research Reports Series, 2015*(2), 1–20. https://doi.org/10.1002/ets2.12076

Sabatini, J., Weeks, J., O'Reilly, T., Bruce, K., Steinberg, J., & Chao, S. (2019). SARA Reading Components Tests, RISE Forms: Technical Adequacy and Test Design, 3rd Edition. *ETS Research Reports Series, 2019*(1), 1–30. https://doi.org/10.1002/ets2.12269

Salsman, J. M., Schalet, B. D., Merluzzi, T. V., Park, C. L., Hahn, E. A., Snyder, M. A., & Cella, D. (2019). Calibration and initial validation of a general self-efficacy item bank and short form for the NIH PROMIS®. *Quality of Life Research, 28*(9), 2513–2523. https://doi.org/10.1007/s11136-019-02198-6

Salsman, J. M., Schalet, B. D., Park, C. L., George, L., Steger, M. F., Hahn, E. A., Snyder, M. A., & Cella, D. (2020). Assessing meaning & purpose in life: Development and validation of an item bank and short forms for the NIH PROMIS®. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation, 29*(8), 2299–2310. https://doi.org/10.1007/s11136-020-02489-3

Sandham, M. H., Medvedev, O. N., Hedgecock, E., Higginson, I. J., & Siegert, R. J. (2019). A Rasch Analysis of the Integrated Palliative Care Outcome Scale. *Journal of Pain and Symptom Management, 57*(2), 290–296. https://doi.org/10.1016/j.jpainsymman.2018.11.019

Sandilos, L. E., Lewis, K., Komaroff, E., Hammer, C. S., Scarpino, S. E., Lopez, L., Rodriguez, B., & Goldstein, B. (2015). Analysis of bilingual children's performance on the English and Spanish

versions of the Woodcock-Muñoz Language Survey-R (WMLS-R). *Language Assessment Quarterly*, *12*(4), 386–408. https://doi.org/10.1080/15434303.2015.1100198

Saracino, R. M., Aytürk, E., Cham, H., Rosenfeld, B., Feuerstahler, L. M., & Nelson, C. J. (2020). Are we accurately evaluating depression in patients with cancer? *Psychological Assessment*, *32*(1), 98–107. https://doi.org/10.1037/pas0000765

Setodji, C. M., Elliott, M. N., Abel, G., Burt, J., Roland, M., & Campbell, J. (2015). Evaluating Differential Item Functioning in the English General Practice Patient Survey: Comparison of South Asian and White British Subgroups. *Medical Care*, *53*(9), 809–817.

Setodji, C. M., Peipert, J. D., & Hays, R. D. (2019). Differential item functioning of the CAHPS® In-Center Hemodialysis Survey. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation*, *28*(11), 3117–3135. https://doi.org/10.1007/s11136-019-02250-5

Shen, Y., Seo, E., Hu, Y., Zhang, M., & Chao, R. K. (2019). Measurement invariance of language brokering extent and attitudes in linguistic minority adolescents: Item response theory analyses. *Cultural Diversity & Ethnic Minority Psychology*, *25*(2), 170–178. https://doi.org/10.1037/cdp0000224

Silverberg, J. I., Lai, J.-S., Vakharia, P. P., Patel, K., Singam, V., Chopra, R., Sacotte, R., Patel, N., Rastogi, S., Kantor, R., Hsu, D. Y., & Cella, D. (2020). Measurement properties of the Patient-Reported Outcomes Measurement Information System Itch Questionnaire item banks in adults with atopic dermatitis. *Journal of the American Academy of Dermatology*, *82*(5), 1174–1180. https://doi.org/10.1016/j.jaad.2019.11.057

Stevanovic, D., Bagheri, Z., Atilola, O., Vostanis, P., Stupar, D., Moreira, P., Franic, T., Davidovic, N., Knez, R., Nikšić, A., Dodig-Ćurković, K., Avicenna, M., Multazam Noor, I., Nussbaum, L.,

Deljkovic, A., Aziz Thabet, A., Petrov, P., Ubalde, D., Monteiro, L. A., & Ribas, R. (2017). Cross-cultural measurement invariance of the Revised Child Anxiety and Depression Scale across 11 world-wide societies. *Epidemiology and Psychiatric Sciences*, *26*(4), 430–440. https://doi.org/10.1017/S204579601600038X

Stone, M. D., Matheson, B. E., Leventhal, A. M., & Boutelle, K. N. (2020). Development and validation of a short form Children's Power of Food Scale. *Appetite*, *147*, 104549. https://doi.org/10.1016/j.appet.2019.104549

Strait, J. E., Wright, E. K. C., & Decker, S. L. (2019). Bender-Gestalt II differential item functioning across Caucasian and African American examinees. *Psychology in the Schools*, *56*(1), 148–158. https://doi.org/10.1002/pits.22181

Studts, C. R., Polaha, J., & van Zyl, M. A. (2017). Identifying unbiased items for screening preschoolers for disruptive behavior problems. *Journal of Pediatric Psychology*, *42*(4), 476–486.

Su, Y., & Behar-Horenstein, L. S. (2018). Assessment of Psychometric Properties of an Oral Health Care Measure of Cultural Competence Among Dental Students Using Rasch Partial Credit Model. *Journal of Dental Education*, *82*(10), 1105–1114. https://doi.org/10.21815/JDE.018.107

Teresi, J. A., Ocepek-Welikson, K., Cook, K. F., Kleinman, M., Ramirez, M., Reid, M. C., & Siu, A. (2016). Measurement equivalence of the Patient Reported Outcomes Measurement Information System® (PROMIS®) Pain Interference short form items: Application to ethnically diverse cancer and palliative care populations. *Psychological Test and Assessment Modeling*, *58*(2), 309–352.

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Ramirez, M., & Kim, G. (2016a). Measurement equivalence of the Patient Reported Outcomes Measurement Information System® (PROMIS®)

Anxiety short forms in ethnically diverse groups. *Psychological Test and Assessment Modeling*, *58*(1), 183–219.

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Ramirez, M., & Kim, G. (2016b). Psychometric properties and performance of the Patient Reported Outcomes Measurement Information System® (PROMIS®) Depression short forms in ethnically diverse groups. *Psychological Test and Assessment Modeling*, *58*(1), 141–181.

Teresi, J. A., Ocepek-Welikson, K., Ramirez, M., Fieo, R., Fulmer, T., & Gurland, B. J. (2018). Development of a Short-Form of the Medication Management Test: Evaluation of Dimensionality, Reliability, Information and Measurement Equivalence Using Latent Variable Models. *Journal of Nursing Measurement*, *26*(3), 483–511. https://doi.org/10.1891/1061-3749.26.3.483

Teresi, J. A., Ocepek-Welikson, K., Ramirez, M., Kleinman, M., Ornstein, K., & Siu, A. (2015). Evaluation of measurement equivalence of the Family Satisfaction with the End-of-Life Care in an ethnically diverse cohort: Tests of differential item functioning. *Palliative Medicine*, *29*(1), 83–96. https://doi.org/10.1177/0269216314545802

Teresi, J. A., Ocepek-Welikson, K., Toner, J. A., Kleinman, M., Ramirez, M., Eimicke, J. P., Gurland, B. J., & Siu, A. (2017). Methodological issues in measuring subjective Well-Being and Quality-of-Life: Applications to assessment of affect in older, chronically and cognitively impaired, ethnically diverse groups using the Feeling Tone Questionnaire. *Applied Research in Quality of Life*, *12*(2), 251–288. https://doi.org/10.1007/s11482-017-9516-9

Terwee, C. B., Crins, M. H. P., Boers, M., de Vet, H. C. W., & Roorda, L. D. (2019). Validation of two PROMIS item banks for measuring social participation in the Dutch general population.

Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation, *28*(1), 211–220. https://doi.org/10.1007/s11136-018-1995-0

Tjipta, S. B., Van De Vijver, F. J. R., Chasiotis, A., & Bender, M. (2019). Contextualized bilingualism among adolescents from four different ethnic groups in Indonesia. *International Journal of Bilingualism*, *23*(6), 1469–1482. https://doi.org/10.1177/1367006918803678

Tucker, C. A., Bevans, K. B., Becker, B. D., Teneralli, R., & Forrest, C. B. (2020). Development of the PROMIS Pediatric Physical Activity Item Banks. *Physical Therapy*, *100*(8), 1393–1410. https://doi.org/10.1093/ptj/pzaa074

Tucker-Seeley, R. D., Marshall, G., & Yang, F. (2016). Hardship among older adults in the HRS: Exploring measurement Differences Across Socio-Demographic Characteristics. *Race and Social Problems*, *8*(3), 222–230. https://doi.org/10.1007/s12552-016-9180-y

Vaingankar, J. A., Abdin, E., Chong, S. A., Sambasivam, R., Jeyagurunathan, A., Seow, E., Picco, L., Pang, S., Lim, S., & Subramaniam, M. (2016). Psychometric properties of the positive mental health instrument among people with mental disorders: A cross-sectional study. *Health & Quality of Life Outcomes*, *14*, 1–13. https://doi.org/10.1186/s12955-016-0424-8

van Amsterdam, J., Vorspan, F., Snijder, M. B., van den Brink, W., Schene, A. H., Stronks, K., Galenkamp, H., & Derks, E. M. (2019). Use of the Fagerström test to assess differences in the degree of nicotine dependence in smokers from five ethnic groups: The HELIUS study. *Drug and Alcohol Dependence*, *194*, 197–204. https://doi.org/10.1016/j.drugalcdep.2018.10.011

van Zyl, M. A., Studts, C., & Pahl, K. (2015). Precision across race, age and gender of a HIV Risk Screen for adolescents and young adults. *Social Work in Public Health*, *30*(3), 260–271. https://doi.org/10.1080/19371918.2014.994725

Vessy, J. A., Strout, T. D., Difazio, R. L., & Ludlow, L. H. (2019). Identifying bullied youth: Re-engineering the Child-Adolescent Bullying Scale into a Brief Screen. *Journal of Applied Measurement*, *20*(4), 367–383.

Ward, J. T., Link, N. W., & Taylor, R. B. (2017). New windows into a broken construct: A multilevel factor analysis and DIF assessment of perceived incivilities. *Journal of Criminal Justice*, *51*, 74–88. https://doi.org/10.1016/j.jcrimjus.2017.06.004

Ward, J. T., Ray, J. V., & Fox, K. A. (2018). Exploring differences in self-control across sex, race, age, education, and language: Considering a bifactor MIMIC model. *Journal of Criminal Justice*, *56*, 29–42. https://doi.org/10.1016/j.jcrimjus.2017.09.006

Wiesner, M., Windle, M., Kanouse, D. E., Elliott, M. N., & Schuster, M. A. (2015). DISC Predictive Scales (DPS): Factor structure and uniform differential item functioning across gender and three racial/ethnic groups for ADHD, conduct disorder, and oppositional defiant disorder symptoms. *Psychological Assessment*, *27*(4), 1324–1336. https://doi.org/10.1037/pas0000101

Xu, R. H., Wong, E. L., Lu, S. Y., Zhou, L., Chang, J., & Wang, D. (2020). Validation of the Toronto Empathy Questionnaire (TEQ) among medical students in China: Analyses using three psychometric methods. *Frontiers in Psychology*, *11*, 810. https://doi.org/10.3389/fpsyg.2020.00810

Yang, C., Ford, M. E., Tilley, B. C., & Greene, R. L. (2016). Religiosity in black and white older Americans: Measure adaptation, psychometric validation, and racial difference. *Medicine*, *95*(37), e4257. https://doi.org/10.1097/MD.0000000000004257

Yang, H., Chen, F., Liu, X., & Xin, T. (2019). An Item Response Theory Analysis of DSM-5 Heroin use disorder in a clinical sample of Chinese adolescents. *Frontiers in Psychology*, *10*, 2209. https://doi.org/10.3389/fpsyg.2019.02209

APPENDIX E: LIST OF ARTICLES WITH DEFINITIONS OF RACE AND/OR ETHNICITY

Definition Provided

Armenta & Cooper (2019); Bowe (2017); Bowe(2019); Dmitrieva et al. (2015); Galenkamp et al. (2017); Galenkamp et al. (2018); Geldenhuys & Bosch (2020); Goetz et al. (2016); Kim et al. (2016); Loyd et al. (2019); Miller et al. (2019); Parkerson et al. (2015); Reich et al. (2018); Rice et al. (2020); Sandham et al. (2019); Sandilos et al. (2015); Terwee et al. (2019); van Amsterdam (2019); van Zyl et al. (2015); Yang et al. (2019)

Proxy for Ethnicity Provided

Ehrich et al (2016); Haroz et al (2016); Jang et al (2018); Lange et al (2016); Roy et al (2016); Stevanovic et al (2017)

No Definition Provided

Abdin et al. (2017); Abdin et al. (2018); Abdin et al. (2019); Adams et al. (2018); Adams et al. (2020); Cartwritght et al. (2020); Chakawa et al. (2015); Charles et al. (2015); Chen & Zhu (2017); Chen et al. (2015); Cicero et al. (2019); Cordier et al. (2019); Crowder et al. (2019); Curran et al. (2018); Devine &Hughes (2016); du Plessis & de Bruin (2015); Dunya et al. (2018); Dunya et al. (2020); DuPaul et al. (2020); Farrington & Lonigan (2015); Fieo et al. (2016); Filshtein et al. (2020); Flanagan (2020); Forrest, Devine et al. (2018); Forrest, Ravens-Sieberer et al. (2018); French & Vo (2020); Gay et al. (2016); Gross et al. (2018); Hahn et al. (2016); Harpole et al. (2015); Harris & Aboujaoude (2016); Harris et al. (2015); Hasnain et al. (2017); Hawes et al. (2018); Heafner & Fitchett (2018); Hong et al. (2016); Hu et al. (2017); Janulis et al. (2018); Jensen et al. (2016); Jones et al. (2016); Kwan et al. (2019a); Kwan et al. (2019b); Lambert, January et al. (2018); Lambert, Garcia et al. (2018); Leung et al. (2020); Li et al. (2020); Lindhiem et al. (2019); Lu et al. (2020); Martin et al. (2020); Mâsse, O'Connor, Lin,

Carbert, et al. (2020); Mâsse, O'Connor, Lin, Hughes, et al. (2020); McDonald et al. (2018); McFarland et al. (2017); Nicholson et al. (2020); Ning (2018); Park et al. (2019); Pascal et al. (2020); Pedersen et al. (2017); Peipert et al. (2018); Quach et al. (2016); Reeve et al. (2016); Rodriguez et al (2019); Rose et al. (2018); Rosselli et al. (2019; Roth et al. (2015); Ruchensky et al. (2020); Ruglass et al. (2020); Salsman et al. (2019); Salsman et al. (2020); Saracino et al. (2020); Setodji et al. (2019); Shen et al. (2019); Silverberg et al. (2020); Stone et al. (2020); Strait et al. (2019); Studts et al. (2017); Su et al. (2018); Teresi et al. (2015); Teresi et al. (2016a); Teresi et al. (2016b); Teresi et al. (2017); Teresi et al. (2018); Teresi, Ocepek-Welikson, Cook et al. (2016); Tjipta (2019); Tucker et al. (2020); Tucker-Seeley et al. (2016); Vaingankar et al. (2016); Vaughn-Coaxum et al. (2016); Vessey et al. (2019); Ward et al. (2017); Ward et al. (2018); Wiesner et al. (2015); Xu et al. (2020); Yang et al. (2016)