8-10-2021

Investigating the Performance of (Multiple-Factor) Multiple-Group Methods for the Detection of Differential Item Functioning

Theresa L. Dell-Ross
Georgia State University

Recommended Citation

Dell-Ross, Theresa L., "Investigating the Performance of (Multiple-Factor) Multiple-Group Methods for the Detection of Differential Item Functioning." Dissertation, Georgia State University, 2021.
doi: https://doi.org/10.57709/24146059

ACCEPTANCE

This dissertation, INVESTIGATING THE PERFORMANCE OF (MULTIPLE-FACTOR) MULTIPLE-GROUP METHODS FOR THE DETECTION OF DIFFERENTIAL ITEM FUNCTIONING, by THERESA L. DELL-ROSS, was prepared under the direction of the candidate's Dissertation Advisory Committee. It is accepted by the committee members in partial fulfillment of the requirements for the degree, Doctor of Philosophy, in the College of Education & Human Development, Georgia State University.

The Dissertation Advisory Committee and the student's Department Chairperson, as representatives of the faculty, certify that this dissertation has met all standards of excellence and scholarship as determined by the faculty.

_____
T. Chris Oshima, Ph.D.
Committee Chair

_____
Kristen L. Buras, Ph.D.
Committee Member

_____
Hongli Li, Ph.D.
Committee Member

_____
Keith D. Wright, Ph.D.
Committee Member

_____
Date

_____
Jennifer Esposito, Ph.D.
Chairperson, Department of Educational Policy Studies

_____
Paul A. Alberto, Ph.D.
Dean, College of Education & Human Development

AUTHOR'S STATEMENT

By presenting this dissertation as a partial fulfillment of the requirements for the advanced degree from Georgia State University, I agree that the library of Georgia State University shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to quote, to copy from, or to publish this dissertation may be granted by the professor under whose direction it was written, by the College of Education & Human Development's Director of Graduate Studies, or by me. Such quoting, copying, or publishing must be solely for scholarly purposes and will not involve potential financial gain. It is understood that any copying from or publication of this dissertation which involves potential financial gain will not be allowed without my written permission.

_____

Theresa L. Dell-Ross

NOTICE TO BORROWERS

All dissertations deposited in the Georgia State University library must be used in accordance with the stipulations prescribed by the author in the preceding statement. The author of this dissertation is:

Theresa L. Dell-Ross
Department of Educational Policy Studies
College of Education & Human Development
Georgia State University

The director of this dissertation is:

T. Chris Oshima, Ph.D.
Department of Educational Policy Studies
College of Education & Human Development
Georgia State University
Atlanta, GA 30303

CURRICULUM VITAE

Theresa L. Dell-Ross

EDUCATION:

Ph.D.	2021	Georgia State University Educational Policy Studies
Ed.S.	2014	Georgia State University Educational Policy Studies
M.Ed.	2003	Vanderbilt University Elementary Education
B.A.	2001	Emory University Educational Studies

PROFESSIONAL EXPERIENCE:

2020-present	Assessment Specialist Georgia Department of Education
2014-2018	Dean's Research Doctoral Fellow College of Education & Human Development Georgia State University
2017-2017	Assessment Design & Evaluation Team Intern Office of Assessment New York City Department of Education
2011-2014	EIP Teacher/Title I School Improvement Specialist Marietta City Schools

PRESENTATIONS AND PUBLICATIONS:

Dell-Ross, T. L., Oshima, T. C., & Wright, K. D. (2017). *Demonstration of Multiple-Factor Multiple-Group Non-Compensatory Differential Item Functioning*. National Council on Measurement in Education Annual Meeting, San Antonio, TX.

Oshima, T. C., & **Dell-Ross, T. L.** (2017). *All Possible Regressions Using IBM SPSS: A Practitioner's Guide to Automatic Linear Modeling*. Georgia Educational Research Association Annual Meeting, Augusta, GA.

PRESENTATIONS AND PUBLICATIONS (CONT.):

Dell-Ross, T. L., Kimball, K. A., & Leroux, A. J. (2016). *Math Performance in the United States: Examining the Effects of Self-Efficacy*. AERA Division D In-Progress Research Gala, Washington, D.C.

Dell-Ross, T. L., & Oshima, T. C. (2015). *Leadership and Governance Patterns in Schools of Native and Immigrant Students Exhibiting High Mathematics Performance*. University Council for Educational Administration Annual Conference, San Diego, CA.

Dell-Ross, T. L. (2014). *Using the Lesson Study Model: One School's Attempt to Increase Elementary Students' Mathematics Achievement and Improve Culture*. Georgia State University Principals Center Symposium, Atlanta, GA. (1st Place Manuscript, Ed.S. Cohort)

PROFESSIONAL SOCIETIES AND ORGANIZATIONS

2014-2018	American Educational Research Association
2014-2018	National Council on Measurement in Education

INVESTIGATING THE PERFORMANCE OF (MULTIPLE-FACTOR) MULTIPLE-GROUP METHODS FOR THE DETECTION OF DIFFERENTIAL ITEM FUNCTIONING

by

THERESA L. DELL-ROSS

Under the Direction of T. Chris Oshima, Ph.D.

ABSTRACT

The examination of assessment items for potential bias is more important than ever. Items that function differently for examinees of equal ability from different groups are said to exhibit differential item functioning (DIF). Traditionally, DIF has been detected by comparing only two groups at a time. In racial/ethnic pairwise comparisons, White examinees were treated as the reference group and one minority group was treated as the focal group. This pairwise analysis was repeated for each minority group of interest. The practice of comparing minority examinees to White examinees must be troubled from a critical race theory perspective. To address the limitations of pairwise analyses, DIF methods that simultaneously analyze items for DIF based on multiple groups and/or multiple grouping factors have been developed. These methods include the generalized Mantel-Haenszel (GMH) statistic and multiple indicators, multiple causes (MIMIC) confirmatory factor analysis (CFA) models. Recently, a multiple-group non-compensatory DIF (MG-NCDIF) index that uses a random sample of all examinees as a base reference group was developed. This study compared the performance of the MG-NCDIF index with the GMH and MIMIC DIF detection methods in simulated conditions that modeled both uniform and non-uniform DIF. Additionally, the GMH and MIMIC methods, which have

historically used a traditional reference group, were modeled using a base group reference. Overall, the MG-NCDIF method exhibited lower power and higher Type I error rates than the MIMIC method. The MG-NCDIF method did outperform the GMH method when non-uniform DIF was simulated via the *a* parameter only; however, when the *b* parameter was manipulated (to model uniform DIF or non-uniform DIF in combination with manipulation of the *a* parameter), power was higher for the GMH index than the MG-NCDIF index. Across analyses, GMH exhibited lower Type I error rates than MG-NCDIF. All three methods exhibited higher power for the detection of uniform DIF and non-uniform DIF when both the *a* and *b* parameters were adjusted; power was lower for the detection of non-uniform DIF when the adjustment was made solely to the *a* parameter. A critical race theory framework guided this study.

INDEX WORDS: Differential item functioning (DIF), Multiple-factor multiple-group non-compensatory DIF, Differential functioning of items and tests (DFIT), Generalized Mantel-Haenszel, Multiple-indicators multiple-causes (MIMIC), Critical race theory, QuantCrit

INVESTIGATING THE PERFORMANCE OF (MULTIPLE-FACTOR) MULTIPLE-GROUP METHODS FOR THE DETECTION OF DIFFERENTIAL ITEM FUNCTIONING

by

Theresa L. Dell-Ross

A Dissertation

Presented in Partial Fulfillment of Requirements for the

Degree of

Doctor of Philosophy

in

Research, Measurement, and Statistics

in

the Department of Educational Policy Studies

in

the College of Education & Human Development

Georgia State University

Atlanta, GA
2021

DEDICATION

This dissertation is dedicated to my husband, Rob, with LOVE (Carla eyes). Each page is the result of your hard work around the house when I was too busy to help, your sense of humor when I needed to lighten up, your hugs when I was too tired and stressed to laugh, and all of the smiles that reminded me what's truly important in life. Here's to the official end of "Boom, boom, boom, another one bites the dust" and taking "another bite out of the elephant"!

LYWAMHMTICSFAFATTPOIAB!

ACKNOWLEDGMENTS

I would like to thank the following people for their insight, guidance, and support during the dissertation process.

Dr. T. Chris Oshima, thank you for years of quantitative instruction, the opportunity to lead, and your unwavering faith. I hope that this paper honors your mentorship.

Dr. Kristen L. Buras, thank you for being the first person to open my eyes and my heart to critical race theory and the marginalized "Other." The conversations that I shared with you have changed the way that I see the world around me. I hope that this paper honors those lessons.

Dr. Hongli Li, thank you for leading me through my first excursions into structural equation modeling and the various software programs used for these analyses. I hope that this paper honors those efforts.

Dr. Keith D. Wright, thank you for collaborating with me on research projects, both past and present. I hope that this paper honors your collegiality.

I am also truly appreciative of the professional kindness afforded by Dr. T. Chris Oshima, Dr. Keith D. Wright, and Dr. Nick White in allowing me to reference their 2015 work and cite their results.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

1 THE PROBLEM

There is currently a "Big Data" movement (Gillborn, Warmington, & Demack, 2018), and assessment is a large part of this movement. Assessment data are used for school and teacher accountability, sometimes in the form of merit pay (Buddin, McCaffrey, Kirby, & Xia, 2007; Hassel & Hassel, 2007), as well as student promotion and retention (for example, Georgia Promotion, Placement, and Retention Law, O.C.G.A. §§ 20-2-282 through 20-2-285). The examination of test items for potential bias is, therefore, more important than ever. Items that function differently for examinees of equal ability from different groups are said to exhibit differential item functioning (DIF). Traditionally, DIF has been detected by comparing only two groups at a time (e.g., Black examinees and White examinees). For researchers and psychometricians, these pairwise comparisons are not ideal for analyzing characteristics of interest that include more than two subgroups (e.g., race/ethnicity, test administration location, native language, test administration language, socio-economic status, academic intervention method, and treatment condition). To address this limitation, DIF methods that simultaneously analyze items for DIF based on multiple groups and/or multiple grouping factors have been developed. These methods include the generalized Mantel-Haenszel (GMH) statistic (Somes, 1986) and multiple indicators, multiple causes (MIMIC) confirmatory factor analysis (CFA) models (Jöreskog & Goldberger, 1975).

DIF analysis is often used to ensure equitable assessment across racial/ethnic subgroups. Historically, in each pairwise comparison, White examinees were treated as the reference group and one minority group was treated as the focal group. This pairwise analysis was repeated for each minority group of interest. Although statisticians could justify the practice of treating White examinees as the reference because their subgroup sample size was largest – a critical property

from a mathematical perspective – this argument is quickly becoming invalid. Based on U.S. Census data, it has been predicted that the number of non-Hispanic White children will *decrease* significantly from 2014 to 2060, from 52% to 35.6%, and that non-Hispanic Whites will represent less than 50% of the U.S. population by the year 2044, making this a "Majority-Minority" nation for the first time (Colby & Ortman, 2015). Furthermore, the practice of comparing minority examinees to White examinees must be troubled from a critical race theory perspective, from which the argument may be made that the use of White examinees as a reference group is reflective of and has contributed to a "Whiteness as the ideal to be reached" mentality. Critical race theorists further argue the inappropriateness of a catch-all minority focal group compared with a White reference group (Gillborn et al., 2018). Another dangerous practice would be the use of particular "model minority" groups as the reference (Gillborn, 2009; Gillborn et al., 2018; Teranishi, 2007).

Recently, a multiple-group non-compensatory DIF (MG-NCDIF) index (Oshima, Wright, & White, 2015) and a multiple-factor multiple-group non-compensatory DIF (MFMG-NCDIF) index (Dell-Ross, Oshima, & Wright, 2017) that use a random sample of all examinees – a base group – as the reference were developed. However, the performance of these measures has yet to be compared with other multiple-group methods. This study was designed to compare the performance of the MG-NCDIF index with the GMH and MIMIC DIF detection methods in simulated conditions that model both uniform and non-uniform DIF. Additionally, the GMH and MIMIC methods, which have historically used a traditional reference group, were modeled using a base group reference. A critical race theory framework guided this study.

Purpose

There were two primary purposes for this study. The first purpose of the study was to investigate the performance of the MG-NCDIF statistic compared with other existing methods that are capable of handling multiple-group analyses. Specifically, MG-NCDIF was compared with GMH and MIMIC indices. The two-group Mantel-Haenszel statistic (Mantel & Haenszel, 1959) has enjoyed widespread popularity (Penfield, 2001); consequently, its multiple-group extension, GMH, is of great interest to researchers. MIMIC models were selected for inclusion because they permit the simultaneous analysis of multiple groups, as well as multiple grouping factors, making them comparable to the MG-NCDIF and MFMG-NCDIF indices. (An additional reason for their inclusion is that the GMH and MIMIC methods have ready-made, user-friendly software options that facilitate automated analyses for test developers and researchers.) The second purpose of the study was to investigate whether the performance of these indices varied by the type of DIF: uniform or non-uniform.

Research Questions

The current study was designed to answer the following questions. First, how does MG-NCDIF perform compared to existing multiple-group DIF detection methods? Second, does the efficacy of the GMH and MIMIC indices vary when detecting various types of DIF (i.e., uniform or non-uniform)?

Significance of the Study

This study is significant for several reasons. First, although a simulation study was conducted to assess the performance of the MG-NCDIF index (Oshima et al., 2015), there has not yet been a simulation study to compare this index with existing methods of DIF detection. Second, most simulation studies that have examined the performance of the GMH and MIMIC

methods did so in the context of uniform DIF; there have been few simulation studies that examined the performance of these methods in the context of non-uniform DIF. Finally, by modeling the MG-NCDIF, GMH, and MIMIC indices using a base (i.e., composite or omnicultural) reference group, the findings of this study contributed to the existing literature base on the use of such a reference group in DIF detection.

2 REVIEW OF THE LITERATURE

Differential Item Functioning

In item response theory (IRT), the probability of an examinee correctly answering an assessment item *i*, denoted as $P_i(\theta)$, is given by the three-parameter logistic (3PL) function

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}},$$

where a_i represents the discrimination parameter for item *i*, b_i is the difficulty parameter for item *i*, c_i is the pseudo-guessing parameter for item *i*, θ represents examinee ability, and *D* is the scaling constant of 1.7. This function can be represented visually; Figure 1 shows this function for two groups of examinees. If there is a difference in probabilities when examinee ability, with respect to the latent trait being measured, is controlled for, then differential item functioning (DIF) is exhibited. There are two types of DIF: uniform and non-uniform. Uniform DIF, indicated by a difference in *b* parameters only, is the case of one group being consistently favored over the other group across the ability continuum. Uniform DIF is shown in Figure 1. As can be seen, with uniform DIF, the item characteristic curves (ICCs) for the two groups do not cross. With non-uniform DIF, on the other hand, one group is favored over the other in the first part of the ability continuum, but then this pattern reverses and the other group is favored in the second part of the ability continuum. In this case, the two ICCs cross, as shown in Figure 2. Non-uniform DIF is the result of groups having either (a) differing *a* parameters or (b) differing *a* and *b* parameters. There are many methods of DIF detection, both within and outside of the IRT framework. For an overview of these methods, readers are referred to Clauser and Mazor (1998) and Magis, Béland, Tuerlinckx, and De Boeck (2010). Regardless of the DIF detection method

selected, DIF may be conceptualized as a performance discrepancy between groups when ability has been taken into account.

To investigate items for DIF, a focal group is compared to a reference group. Historically, the reference group, arbitrarily selected by the researcher, has been the "majority" (i.e., the subgroup with the highest rate of participation), because parametric estimations, such as those conducted in IRT-based analyses, are more stable with larger sample sizes. In the case of racial/ethnic analyses, White examinees have typically been used as the reference. The focal group has traditionally been a select minority subgroup. Such "pairwise" comparisons have been the tradition for DIF detection (Fidalgo & Scalon, 2010; Magis et al., 2010); they have been the go-to for assessment companies for decades.

With pairwise DIF detection, multiple analyses are conducted for a single item. For example, if the researcher is concerned that an item may exhibit DIF for Asian, Black, or Hispanic examinees, three separate tests are conducted: Asians compared to Whites, Blacks compared to Whites, and Hispanics compared to Whites. There are three main problems with such pairwise DIF analysis. First, the multiple pairwise tests require an adjustment of the alpha level to avoid inflated Type I error rates. Second, these pairwise comparisons have lower power for DIF detection than simultaneous multiple-group testing methods (Magis et al., 2010). Third, as discussed at more length later in this paper, the arbitrary selection of a reference group is problematic; changing demographics and new perspectives on the way quantitative research contributes to systematic and institutionalized racism call for alternative means of DIF detection.

Critical Race Theory

QuantCrit.

Aimed at exposing systematic, institutionalized racism and fostering social justice by contextualizing human experience and giving a voice to marginalized groups, critical race theory (CRT) has historically been the purview of qualitative research. More recently, quantitative researchers with an interest in social justice and critical race theorists with an interest in quantitative research have begun to break down the wall between CRT and statistics, looking to CRT as a guiding framework. This methodological crossover with its numerous distinctions has come to be known by various names, including *Critical Race Transformative Convergent Mixed Methods* (Garcia & Mayorga, 2018), *Critical Race Quantitative Intersectionality* (Covarrubias et al., 2018), *CritQuant* (Sullivan, Larke, & Webb-Hasan, 2010), and – to be used as a framework for this paper – *QuantCrit* (Gillborn et al., 2018).

To explicate the ideals of QuantCrit, an understanding of the main principles of CRT is imperative. First and foremost, CRT scholars believe that "racism is prevalent in all aspects of society, with schools not being an exception" and that this "notion of the permanence of racism suggests that racist hierarchical structures govern all political, economic, and social domains" (DeCuir & Dixson, 2004, pp. 26 & 27, respectively). As Ladson-Billings and Tate (2006) explained, "when we speak of racism we refer to Wellman's definition: 'culturally sanctioned beliefs which, regardless of the intentions involved, defend the advantages Whites have because of the subordinated positions of racial minorities'" (pp. 18-19). Racism, as a pervasive force in our society, has permeated all our institutions, and "not merely in those spaces seen as racially defined spaces" (Ladson-Billings, 2004, p. 5). Thus, it influences education and the academy, teachers and researchers. It follows logically therefore that "critical race methodology in

education challenges White privilege, rejects notions of 'neutral' research or 'objective' researchers, and exposes deficit-informed research that silences or distorts epistemologies of people of color" (Solórzano & Yosso, 2009, p. 133).

Consequently, critical race theorists call for research that is situated, both historically and contextually, and that recognizes the voices, perspectives, and experiential knowledge of marginalized and oppressed peoples (see, for example, Buras, 2014; Ladson-Billings, 2009; Morris, 2006). As Pérez Huber, Vélez, and Solórzano (2018) explained, "numbers can offer vital insights, highlight patterns, and convey particular analyses, but when they are decontextualized, ahistorical, and disconnected from the everyday lives of People of Color, *they are hypothetical at best*" (p. 212). Covarrubias et al. (2018) further pointed out the dangers of numbers in isolation: "Without sociohistorical contexts, these interpretations run the risk of perpetuating deficit ideologies about the causes that produce and reproduce these outcomes" (p. 253). This is especially true in the arena of educational research, where the "achievement gap" is a common element in the dominant discourse. Covarrubias et al. (2018) cautioned that "educational pipelines require critical pedagogies to situate and deconstruct static numbers if we are to capture the complexity of lived experiences and challenge stereotypes of a monolithic educational trajectory for entire Communities of Color." (p. 253). Similarly, Buras (2014) argued that high-quality research involves "race-conscious ethics and a long-term relationship with the community" and "cannot be done along well-established lines that call for aloofness, distance, objectivity, color-blindness, neutrality, and the 'untainted' judgments of the all-knowing researcher about, not with, the 'objects' of study" (p. 35). Stories – and counterstories – "add necessary contextual contours to the seeming 'objectivity' of positivist perspectives" (Ladson-Billings, 2009, pp. 21-22).

Researchers studying race – whether qualitatively or quantitatively – would do well to adopt a CRT perspective. Race is a social construct, not an objective, scientific, biological characteristic as it is often treated (Ladson-Billings, 2009). Furthermore, researchers often neglect to recognize that "objective" quantitative research and statistical analyses are themselves as much a social construction as race (Gillborn, 2010). Quantitative research is perceived to be objective, neutral, and scientific even though it is a social construction, susceptible to the researcher's biases and experiences (Gillborn, 2010). As Morris (2006) asserted, "the race, social class or political views of the researcher affect the research process, because researchers bring their own epistemological perspectives – ways of knowing – into the framing of researchable questions, data collection and analysis, and interpretations and conclusions" (p. 133). From a CRT perspective, there is no such entity as "objective research"; it reflects the subjectivities of the researcher. There is no point in the research process in which researchers' biases and prejudices are isolated from the research itself; indeed,

> cultural influences have set up the assumptions about the mind, the body, and the universe with which we begin; pose the questions we ask; influence the facts we seek; determine the interpretation we give these facts; and direct our reaction to these interpretations and conclusions. (Gould, 1996, p. 55)

Similarly, Crenshaw, Gotanda, Peller, and Thomas (1995) argued, "there is 'no exit' – no scholarly perch outside the social dynamics of racial power from which merely to observe and analyze. Scholarship – the formal production, identification, and organization of what will be called 'knowledge' – is inevitably political" (p. xiii).

Gillborn at al. (2018) made a similar point. In discussing the public's surprised response to the racially-biased results of a computer algorithm, they explained

> we argue that, far from being surprised that quantitative calculations can reproduce human bias and racist stereotypes, such patterns are entirely predictable and should lead us to treat quantitative analyses with at least as much caution as when considering qualitative research and its findings . . . . Simply because the mechanics of an analysis are performed by a machine does not mean that any biases are automatically stripped from the calculations. On the contrary, not only can computer-generated quantitative analyses embody human biases, such as racism, they also represent the *added* danger that their assumed objectivity can give the biases enhanced respectability and persuasiveness. (p. 159)

They rightly pointed out the weight that quantitative data carry in our society. Gould (1996) had already made the same argument, stating that a "reason for analyzing quantitative data arises from the special status that numbers enjoy. The mystique of science proclaims that numbers are the ultimate test of objectivity" (p. 58). This "special status" can be seen clearly in the "Big Data" movement: "Big Data has become big business . . . . where, most significantly, theories and human reasoning are rendered obsolete because the 'numbers speak for themselves'" (Gillborn et al., 2018, pp. 165-167). Quantitative research has long enjoyed an elevated status above qualitative research and thus contributes to enduring systematic marginalization of people of color. Contrary to this position, CRT scholars have a "preference of the experiences of oppressed peoples (narrative) over the 'objective' opinions of whites" (Taylor, 2009, p. 4). Research must, therefore, be situated historically and contextually, giving voice to those who have heretofore been rendered mute. Unfortunately, this rarely occurs in quantitative research.

Gillborn et al. (2018) defined QuantCrit not as a new theory, but as an extension of CRT; it is the practice of quantitative research from a CRT perspective. They outlined five guiding

principles “as a kind of toolkit that embodies the need to apply CRT understandings and insights whenever quantitative data is used in research and/or encountered in policy and practice” (p. 169). Garcia, López, and Vélez (2018) summarized these five principles:

> (1) The centrality of racism as a complex and deeply rooted aspect of society that is not readily amenable to quantification; (2) The acknowledgement that numbers are not neutral and they should be interrogated for their role in promoting deficit analyses that serve white racial interests; (3) The reality that categories are neither ‘natural’ nor given and so the units and forms of analysis must be critically evaluated; (4) The recognition that voice and insight are vital: data cannot ‘speak for itself’ and critical analyses should be informed by the experiential knowledge of marginalized groups; (5) The understanding that statistical analyses have no inherent value but they can play a role in struggles for social justice. (p. 151)

QuantCrit scholars acknowledge the limitations of statistical analyses: “every attempt to ‘measure’ the social in relation to ‘race’ can only offer a crude approximation that risks fundamentally misunderstanding and misrepresenting the true nature of the social dynamics that are at play” (Gillborn et al., 2018, p. 170). In “promoting deficit analyses that serve white racial interests,” quantitative analyses function as “racial projects.” Racial projects, according to Omi and Winant (1994) are “efforts to shape the ways in which human identities and social structures are racially signified, and the reciprocal ways that racial meaning becomes embedded in social structures” (p. 13).

Although “statistics, studies, and formal databases, while important, do not, in the view of CRT scholarship, have the moral certitude, proficiency, and knowledge base adequate to name and resist oppression” (Taylor, 2009, p. 5), QuantCrit scholars such as Gillborn (2010) do not

advocate the abandonment of statistical analysis, but instead recommend "using the master's tools" (p. 270). As Gillborn explained, "the fact that stats have such force in the public consciousness means that we have to be more imaginative – and critical – in how we use them to fight racism" (p. 271). As a social construct, statistical analysis is neither positive nor negative by definition; it is the implementation and use of such research that ultimately renders it as a positive or negative force in the push for racial/ethnic equity; "with appropriate safeguards and reflexivity quantitative material has the potential to contribute to a radical project for greater equity in education" (Gillborn et al., 2018, p. 160). Thus, a racial project may become a racial justice project.

Positionality.

Quantitative researchers have been criticized for their failure to acknowledge their own positionalities, opting instead to present "an ice-cold impartiality" with "dispassionate objectivity," even though "impartiality (even if desirable) is unattainable by human beings with inevitable backgrounds, needs, beliefs and desires" (Gould, 1996, p. 36). Gould recognized the peril of believing oneself to be neutral "for then one stops being vigilant about personal preferences and their influences – and then one truly falls victim to the dictates of prejudice" (p. 36). This is CRT at its core. Instead of representing an unattainable ideal as the definition of objectivity, Gould proposed that it be reimagined as "fair treatment of data, not absence of preference" (p. 36), explaining that "the best form of objectivity lies in explicitly identifying preferences so that their influence can be recognized and countermanded" (p. 37).

For Taylor (2009), "positionality then becomes a perspective that must be disclosed; it identifies the frame of reference from which researchers, practitioners, and policy makers present their data, interpretations, and analysis" (p. 8). However, with the significance that has been

assigned to quantitative simulation studies, his statement should perhaps be modified to read "positionality . . . identifies the frame of reference from which researchers, practitioners, and policy makers present their *analytical design*, data, interpretations, and analysis." Gould (1996) and Taylor (2009) are not alone in calling for self-reflexivity and disclosure on the part of the researcher; indeed, the inclusion of a positionality statement in empirical research is beginning to take hold with others (see, for example, Covarrubias et al., 2018; Garcia & Mayorga, 2018; López, Erwin, Binder, & Chavez, 2018).

It is here that I would like to disclose my own positionality, although using the personal pronoun "I" is anathema to me as a quantitative researcher. Any scholar who has read quantitative research extensively is well aware that personal pronouns are eschewed; if one is desperate, one may refer to oneself as "one" or "the author" or a similar referent. In an effort to honor both the CRT and quantitative research traditions, I have decided to include a positionality statement. However, I beg the reader's forgiveness when I revert to third-person language in subsequent sections as my training dictates. I shall rejoin you in this more personal manner in the Discussion chapter.

My social location is a privileged one by any classification; I am a White, middle-class, middle-aged, abled, heterosexual woman who has enjoyed the benefits of being a natural-born citizen of the U.S. and a native English speaker, with the opportunity to pursue and earn advanced educational degrees. After 11 years as a teacher in the public-school system, I embarked on a new professional journey to earn a Ph.D. in research, measurement, and statistics. Specifically, my primary interest is the intersection of statistics and assessment, known as "measurement" or "psychometrics." I have never viewed myself – nor do I now – as a social justice activist. I am simply looking to, as doctors say, "first, do no harm." I do not want my

work to contribute to the marginalization or subordination of others; indeed, this is what motivates my interest in the identification of test item bias.

It was my goal to abide by the tenets of CRT and QuantCrit in this study, insofar as I was able, given that simulation studies represent the most decontextualized form of quantitative research. From the CRT and QuantCrit perspectives, there is a need to trouble the perceived objectivity and truthfulness of data and statistics, and I wished to honor that goal. To my mind, the best way for me to achieve this goal in the current paper was to identify the subjective (i.e., arbitrary) decisions that I was required to make throughout this research, acknowledging that an entirely different set of results may have been reached if I were to have made different choices. Doing so, I believe, is a first step for quantitative researchers who are concerned with the potential (ab)uses to which their work might be put. After all, as Gillborn et al. (2018) cautioned,

> statistics are socially constructed in exactly the same way that interview data and survey returns are constructed i.e. through a design process that includes, for example, decisions about which issues should (and should not) be researched, what kinds of questions should be asked, how information is to be analyzed, and which findings should be shared publicly . . . . at every stage there is the possibility for decisions to be taken that obscure or misrepresent issues that *could* be vital to those concerned with social justice. (p. 163)

DIF detection and the traditional White reference group.

The inclusion of QuantCrit principles is a noble goal in its own right, regardless of the quantitative study at hand. Doing so seems even more imperative given the nature of the topic at hand: multiple-group DIF. One of the most significant uses of DIF analysis is to ensure

racial/ethnic fairness in large-scale assessment programs. As already discussed, White examinees have traditionally been used as the reference group in pairwise DIF comparisons.

From a CRT perspective, there is certainly room to argue the treatment of White data in this way was reflective of and contributed to a "Whiteness as the ideal to be reached" mentality. Taylor (2009) summed it up succinctly: "White supremacy is the background against which other systems are defined" (p. 4). His statement could easily be modified to "White students' academic performance is the background against which other students' performance is defined."

This mentality played a role in school desegregation. Policymakers assumed that White students, teachers, and schools were more successful. Indeed, Morris (2006), in his interviews with Black educators involved in school desegregation in the mid-1990s, found a theme of "the stigmatizing of black teachers as incompetent and the subsequent stigmatizing of all-black schools as 'inferior' institutions" (p. 136). Thus, placing Black students in White schools was believed to be the "fix," as Solórzano and Yosso (2009) explained: "The main solution for the socioacademic failure offered by cultural deficit majoritarian storytellers is cultural assimilation. Specifically, they argue that students of color should assimilate to the dominant White middle-class culture to succeed in school and in life" (p. 138). However, as one of the principals in Morris' (2006) study pointed out, "I've never been a proponent of sending black children to sit with white children was going to help them to learn" (p. 141).

Although the use of a White reference group has been criticized by some critical race theory scholars (e.g., Garcia & Mayorga, 2018), it should be mentioned that to choose, instead, particular minority groups to serve as the reference group is no less dangerous. This is because

> certain groups are held up as 'model minorities', a stereotype of hard work and success that harms both the group itself (by obscuring certain other disadvantages,

> such as higher rates of unemployment) and, by implication, other less successful groups (whose 'failure', it is reasoned, must surely be their own fault). (Gillborn, 2009, pp. 60-61)

The use of model minorities as the reference obscures intragroup heterogeneity. For example, in his discussion regarding the umbrella label "Asian American," Teranishi (2007) argued that "the perceived educational success of Asian Americans has resulted in their exclusion altogether from racial discourse on educational issues because it is believed that there is no need to address their educational issues" (pp. 47). Critical race theorists further assert that the creation and use of a catch-all minority focal group compared with a White reference group is also inappropriate (Gillborn et al., 2018).

Regardless of which group is selected as the reference, it is a subjective decision made by the researcher. Therefore, the type of reference group required for different DIF detection methods must be considered very carefully when weighing the costs and benefits of various indices.

Multiple-Factor Multiple-Group Non-Compensatory DIF

As stated earlier, there are a plethora of methods for the detection of DIF in test items. In the differential functioning of items and tests (DFIT) framework (Raju, van der Linden, & Fleer, 1995), one method of measuring DIF is the non-compensatory DIF (NCDIF) index. The NCDIF index for dichotomous items is defined as the expected value of the squared distance between the probabilities of a correct response for the reference group and the focal group at a given theta (ability). Specifically, NCDIF for item *i* is calculated as

$$\text{NCDIF}_i = \frac{\sum_{s=1}^{N_F} d_i(\hat{\theta}_s)^2}{N_F},$$

where $d_i\left(\hat{\theta}_s\right)^2$ is the squared difference between $P_{iF}(\hat{\theta})$, the probability of correct response for examinee *s* on item *i* at a given $\hat{\theta}$ using the item parameter estimates from the focal group, and $P_{iR}(\hat{\theta})$, the probability of correct response for examinee *s* on item *i* at the same given $\hat{\theta}$ using the item parameter estimates from the reference group. N_F is the sample size of the focal group. The squaring of the difference between probabilities is a critical element of the calculations, as it prevents differences that favor the reference group and differences that favor the focal group from cancelling each other out. In other words, by squaring the differences between the reference group's ICC and the focal group's ICC, both uniform and non-uniform DIF may be detected (Oshima & Morris, 2008).

The NCDIF statistic may be tested for significance via the item parameter replication (IPR) method (Oshima, Raju, & Nanda, 2006). In an IPR test of significance, for each item, the focal group's item parameters and variance-covariance estimates are used to create a set of simulated item parameters with the same variance and covariance structure. These simulated item parameters are then randomly paired, as though one represented the estimations for a reference group and one represented the estimations for a focal group. Next, NCDIF is calculated for each pair of parameters and rank ordered. As the simulated parameters come from the same distribution, all pairwise differences between the two simulated groups are due to sampling error (Oshima & Morris, 2008; Oshima et al., 2006). Any observed NCDIF value beyond the $(1 - \alpha)$ rank is deemed extreme and, therefore, significant. This process has been automated in the "DIFCUT" program (Nanda, Oshima, & Gagne, 2005).

Oshima et al. (2015) extended this pairwise NCDIF index to a multiple-group NCDIF (MG-NCDIF) index. MG-NCDIF is defined as

$$\text{MG-NCDIF}_i = \frac{\sum_{g=1}^{p}\sum_{s=1}^{N_B} d_{ig}\left(\hat{\theta}_s\right)^2}{pN_B},$$

where $d_{ig}(\hat{\theta}_s) = P_{iB}(\hat{\theta}_s) - P_{iG_g}(\hat{\theta}_s)$ for group *g* of *p* groups. In moving from a pairwise comparison to a multiple-group comparison, an arbitrary reference group is no longer selected. Instead, MG-NCDIF utilizes a "base" group comprised of a random sample (or a stratified random sample) of all examinees. MG-NCDIF results may be tested for significance with the same IPR method discussed above. As significance tests are sensitive to sample size, Oshima et al. (2015) recommended that the size of the base group be equal to the average subgroup sample size from the complete examinee dataset. N_B is the sample size for this base group, *B*. The base group data is then used in the IPR method to determine the cutoff value for significance for each item.

Dell-Ross, Oshima, and Wright (2017) further extended MG-NCDIF to include multiple grouping factors (i.e., variables such as gender and race/ethnicity). Multiple-factor multiple-group NCDIF (MFMG-NCDIF) is given as

$$\text{MFMG-NCDIF}_i = \frac{1}{q}\sum_{k=1}^{q}\left[\frac{\sum_{g=1}^{p}\sum_{s=1}^{N_B} d_{igk}\left(\hat{\theta}_s\right)^2}{pN_B}\right],$$

where MG-NCDIF$_i$ is averaged for *q* factors, making it essentially the average of the squared *d* values, unweighted by the number of groups within each factor. By calculating MG-NCDIF for each factor and then averaging these values, each factor is given equal weight. Thus, a factor with many levels is not favored over a factor with only a few levels. [For example, racial/ethnic DIF (several levels) would be weighted equally with gender DIF (two levels).] It should be noted that a researcher could, if interested, weight the factors. However, this would be an arbitrary

decision, and is not recommended unless there is an imperative reason for doing so. Again, the base group data is used to determine significance of the MFMG-NCDIF results.

The use of these NCDIF indices has several advantages. First, as just described, these relatively new indices may be tested for significance. Second, the DFIT framework includes an index for differential test functioning (DTF), which is analogous to testing for DIF across an entire assessment. Just as NCDIF was extended to test for multiple-factor multiple-group DIF detection, DTF has been extended to multiple-factor multiple-group DTF detection (Dell-Ross et al., 2017). Third, the NCDIF indices are weighted by density, or the number of examinees at each value of theta, such that the observations at the extreme ends of the ability continuum do not exert undue influence in the analysis. Fourth, for each item, MG-NCDIF and MFMG-NCDIF return a single result. There is no need for multiple testing of a single item and, therefore, there is no need to adjust the alpha level for multiple tests. Last – and most importantly – as the MG-NCDIF and MFMG-NCDIF statistics use a base group, researchers are not required to arbitrarily select a reference group. Therefore, comparisons are not a matter of social convention or majoritarian sample size.

There is one potential disadvantage to these NCDIF indices. As the amounts of DIF are averaged, it is possible that the addition of groups and/or grouping factors may obscure DIF. To date, there has been only one simulation study to assess the performance of the MG-NCDIF index: Oshima et al. (2015). Oshima et al. simulated a variety of conditions to assess the efficacy of the new MG-NCDIF statistic, including sample size, type of DIF, DIF pattern, and impact, across three- and five-group conditions. They found that the performance of the MG-NCDIF index was not affected by the number of groups and that this new DIF index had Type I error rates and power comparable to the two-group (pairwise) NCDIF analysis. They also concluded

that the MG-NCDIF statistic accurately detected both uniform and non-uniform DIF and that the direction of impact (unidirectional or bidirectional) did not affect the results, both of which are promising findings for this new test of DIF. To date, there have not been any simulation studies to assess the efficacy of the MFMG-NCDIF index, so it remains unclear if the addition of factors, which will likely increase the total number of groups, will lead to a "watered down" DIF identification rate.

Generalized Mantel-Haenszel

DIF between two groups can also be detected using the Mantel-Haenszel (MH) method (Mantel & Haenszel, 1959). For dichotomously-scored items, MH is one of the most popular DIF detection methods (Fidalgo, 2011; Magis et al., 2010; Penfield, 2001). In the MH method, a 2×2 contingency table is used to test the relationship between group membership and item response while controlling for total test score (the sum, or matching, score). The null hypothesis is that there is no association between group membership and item response; the response variable is distributed randomly. In the MH framework, dichotomous items are investigated for DIF using the formula

$$\mathrm{MH} = \frac{\left(\left|\sum_j A_j - \sum_j \mathrm{E}(A_j)\right| - 0.5\right)^2}{\sum_j \mathrm{Var}(A_j)},$$

where A_j represents the number of correct responses among reference group examinees with sum score j. E(A_j) and Var(A_j) are calculated, respectively, as

$$\mathrm{E}(A_j) = \frac{n_{Rj} m_{1j}}{T_j}$$

and

$$\mathrm{Var}(A_j) = \frac{n_{Rj} n_{Fj} m_{1j} m_{0j}}{T_j^2 (T_j - 1)},$$

where n_{Rj} is the number of responses among reference group examinees with sum score *j*, n_{Fj} is the number of responses among focal group examinees with sum score *j*, m_{1j} is the number of correct responses among examinees with sum score *j*, and m_{0j} is the number of incorrect responses among examinees with sum score *j*. T_j is the number of examinees with sum score *j*. MH follows a χ^2 distribution with one degree of freedom.

MH was extended to detect DIF simultaneously in more than two groups using the GMH index (Landis, Heyman, & Koch, 1978; Penfield, 2001; Somes, 1986). GMH is given as

$$\mathrm{GMH} = \left(\sum_j \boldsymbol{A}_j - \sum_j \mathbf{E}(\boldsymbol{A}_j) \right)' \left(\sum_j \mathbf{Var}(\boldsymbol{A}_j) \right)^{-1} \left(\sum_j \boldsymbol{A}_j - \sum_j \mathbf{E}(\boldsymbol{A}_j) \right),$$

where all variables are defined as above, except that $\mathbf{E}(\boldsymbol{A}_j)$ and $\mathbf{Var}(\boldsymbol{A}_j)$ are now vectors. For dichotomous items, GMH follows a χ^2 distribution with $G - 1$ degrees of freedom, where G is the number of groups. GMH conducts a single test of significance, testing the null hypothesis that DIF is observed in at least one pair of groups. If the source of DIF is of interest to the researcher, post-hoc pairwise analyses may be conducted using the MH method (Finch, 2016) or MH with a Bonferroni-adjusted alpha level (Penfield, 2001). GMH is a specific case of the generalized nominal Mantel-Haenszel $Q_{GMH(1)}$ statistic (Fidalgo & Scalon, 2010).

The GMH index has several advantages over other DIF detection methods. First, due to the popularity of the MH class of methods, there are two pieces of software that are ready-made for multiple-group GMH analyses: GMHDIF (Fidalgo, 2011) and the "difR" package (Magis et al., 2010). Second, as just mentioned, GMH simplifies DIF detection by providing a single index across multiple groups (Penfield, 2001), alleviating the need for researchers to make arbitrary

decisions on how to proceed when DIF is observed between some groups but not others; this also simplifies the purification process. Third, GMH does not conduct the estimations that parametric methods such as MG-NCDIF and MIMIC models undergo. This means that GMH analyses may be reliably conducted with much smaller sample sizes (Fidalgo & Madeira, 2008), a characteristic that is critical as many focal groups do not meet the minimum sample size guidelines that parametric methods require (Fidalgo & Scalon, 2010). A related advantage is that GMH does not require a specific parametric function to be fit to the data (Fidalgo & Scalon, 2010; Wang & Su, 2004). Fourth, Type I error rates are well-controlled with GMH, as multiple pairwise tests are avoided (Fidalgo & Scalon, 2010; Penfield, 2001). This is important because, even when the nominal alpha level is controlled, power is lower for multiple pairwise tests (Penfield, 2001). Finally, the MH family of methods has been shown to have high power for detecting uniform DIF (Fidalgo & Madeira, 2008; Finch, 2016; Sireci & Rios, 2013).

The GMH statistic also has a few disadvantages in regard to DIF detection. First, MH and GMH are much less sensitive to the detection of non-uniform DIF than other DIF detection procedures (Penfield, 2001; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990; Wang & Su, 2004); in other words, GMH performs best when the data are fit to a 1PL model (Fidalgo & Madeira, 2008; Penfield, 2001). Second, a reference group must still be arbitrarily selected by the researcher. For example, if a researcher wished to study items for potential bias across White, Asian, Black, and Hispanic examinees, then Asian examinees would be compared with White examinees, Black examinees would be compared with White examinees, and Hispanic examinees would be compared with White examinees (assuming White examinees were treated as the reference group), despite the fact that this is an omnibus test. Third, a GMH effect size has not yet been developed. Therefore, in cases where a significant finding may be an artifact of

large sample sizes, there is no means of analyzing whether this is indeed the case (Penfield, 2001).

Despite these drawbacks to the GMH index, there is some evidence that it may be a powerful tool for multiple-group DIF detection. Fidalgo and Scalon (2010) found that the $Q_{GMH(1)}$ statistic (of which GMH is a specific case) and Bonferroni-adjusted GMH index had acceptable Type I error rates (ranging from .048 to .067 and .046 to.065, respectively) across conditions with two or three focal groups (three or four groups total). They also concluded that observed power levels for these two indices were acceptable, again across conditions with two or three focal groups. In a multiple-group comparison of GMH, generalized logistic regression (Magis, Raîche, Béland, & Gérard, 2011), Lord's chi-square test (Lord, 1980), and the multiple-group alignment procedure (Asparouhov & Muthén, 2014), Finch (2016) found that GMH had the second-best Type I error rates, was less sensitive to the number of groups analyzed when sample sizes were equal, exhibited excellent power across sample sizes (N = 500, 1000, and 2000) when DIF magnitude was 0.6 or 0.8, and exhibited excellent power in unequal sample size conditions regardless of the number of groups. He concluded that GMH and the alignment method had the best combination of Type I error rates and power when there were more than two groups in the analysis. Penfield (2001) compared GMH, MH, and a Bonferroni-adjusted MH (BMH). Although GMH Type I error rates were slightly higher than those of BMH, these rates were still within the nominal alpha of .05. He also found that GMH and MH consistently exhibited the highest power levels. As GMH had the best balance between Type I error rates and power, he recommended the use of GMH over MH or BMH in multiple-group analyses.

Multiple Indicators, Multiple Causes Confirmatory Factor Analysis

More recently, DIF has been detected using multiple indicators, multiple causes (MIMIC) confirmatory factor analysis (CFA) models (Jöreskog & Goldberger, 1975; see, for example, Chun, Stark, Kim, & Chernyshenko, 2016; Finch, 2005; Shih & Wang, 2009; Wang & Shih, 2010; Woods & Grimm, 2011; Woods, Oltmanns, & Turkheimer, 2009). In a MIMIC model, a direct path extends from the latent factor to each assessment item. (Technically, this path extends from the latent factor to the latent response variable for each item, which in turn extends to the observed response variable for each item; however, the latent response variables are often omitted from the path diagrams for parsimony.) Additionally, a direct path extends from the grouping factor(s) to the latent factor. To test an item for uniform DIF, direct paths are added from the grouping factor(s) to the item being studied. Figure 3 presents a visual representation of this MIMIC model. Mathematically, in relation to IRT, this MIMIC model takes the form

$$y_i^* = \lambda_i \theta + \beta_{ij}' z_j + \varepsilon_i$$

and

$$y_i = \begin{cases} 1 \text{ if } y_i^* \geq \tau_i \\ 0 \text{ if } y_i^* < \tau_i \end{cases},$$

where y_i^* is the latent response variable *i*, λ_i is the factor loading for variable *i*, θ is the latent trait, z_j is the dummy variable indicating group *j* membership, β_{ij}' is the direct loading from group *j* to item *i*, ε_i represents the error associated with item *i*, y_i is the observed dichotomous response for item *i*, and τ_i represents the threshold for item *i*. The threshold parameter τ and the factor loading λ are related to IRT item difficulty and/or discrimination in a two-parameter logistic (2PL) model as follows:

$$\lambda_i = \frac{a_i / D}{\sqrt{1+(a_i / D)^2}}$$

and

$$\tau_i = \frac{(a_i / D)b_i}{\sqrt{1+(a_i / D)^2}},$$

where a_i is the discrimination of item *i*, b_i is the difficulty of item *i*, and *D* is the scaling constant of 1.7. A significant β'_{ij} value indicates the presence of uniform DIF. It should be noted that, although MIMIC models are most commonly referred to as CFA models, they are actually structural regression models, due to the endogenous nature of the MIMIC factor (Kline, 2016). They have been labeled as CFA models herein to provide consistency with existing literature.

Although MIMIC was originally designed for the detection of uniform DIF, it has recently been extended to detect non-uniform DIF (Woods & Grimm, 2011; see, also, Chun et al., 2016). To detect non-uniform DIF, an interaction between the latent trait and each dummy-coded grouping variable is added to the model, with a path extending from each interaction term to the studied item, as shown in Figure 4. The resulting model is

$$y_i^* = \lambda_i\theta + \beta'_{ij}z_j + \omega_{ij}\theta z_j + \varepsilon_i,$$

where $\omega_{ij}\theta z_j$ represents this interaction and all other variables are as defined previously. Interestingly, very few studies have been conducted investigating the performance of MIMIC models when non-uniform DIF is present. In fact, some authors continue to erroneously state that MIMIC models may not be used to detect this type of DIF (e.g., Pendergast, von der Embse, Kilgus, & Eklund, 2017).

The dearth of simulation studies in this area may be related to the complexity of estimating this model. Unfortunately, the calculation of an interaction term involving a latent

trait is not a straightforward multiplicative matter. Instead, Klein and Moosbrugger's (2000) latent moderated structural (LMS) equation algorithm is employed to estimate these terms (Finch & French, 2019; Woods & Grimm, 2011). The LMS algorithm, first developed to estimate non-linear effects in CFA models, can also be applied to interaction terms involving latent variables. The M*plus* program (Muthén & Muthén, 2007), which appears to be the most commonly used software for MIMIC analyses, features the "XWITH" command for the creation of interaction terms involving latent traits. Analyses conducted with XWITH utilize the LMS algorithm and robust maximum likelihood (MLR) estimation with numerical integration (Muthén & Muthén, 2017). [For a detailed treatment of numerical integration, which is beyond the scope of this paper, readers are referred to Muthén (2004).] It should be noted that the LMS method assumes a normal distribution for each variable included in the interaction term and that this assumption is violated in the case of DIF analyses where one of the variables is a dummy-coded grouping variable; consequently, an inflation of the Type I error rate may be observed (Klein & Moosbrugger, 2000; Woods & Grimm, 2011), even as the power for detection of non-uniform DIF remains sufficient (Finch & French, 2019).

To conduct two-stage omnibus testing for uniform and/or non-uniform DIF, Lopez Rivas, Stark, and Chernyshenko (2009) and Chun et al. (2016) recommended a constrained baseline approach at Stage 1 to identify anchor items and a free baseline approach at Stage 2 to identify items that function differentially. In the constrained baseline approach (Stage 1), a more restrictive model [Figure 5(a)] consisting of paths from the grouping variables to the latent trait, γ_j, and from the latent trait to each item, λ_i, is compared with a less restrictive model [Figure 5(b)] in which (1) paths from the grouping variables to the studied item, β_{ij}, have been added to test for uniform DIF and (2) the interaction terms, θz_j, and paths from these terms to the studied

item, ω_{ij}, have been added to test for non-uniform DIF. This model comparison is conducted separately for each item; a significant difference between models indicates that the item exhibits DIF. Thus, items in which the difference test is non-significant are eligible for use as anchors in Stage 2.

In the free baseline approach (Stage 2), researchers begin with a less restrictive model [Figure 6(a)] consisting of paths from (1) the grouping variables to the latent trait, γ_j, (2) the latent trait to each item, λ_i, (3) the grouping variables to all non-anchor items, β_{ij}, and (4) the interaction terms to all non-anchor items, ω_{ij}. This model is compared with a more restrictive model [Figure 6(b)] in which the paths from the grouping variables to the studied item, β_{ij}, and the paths from the interaction terms to the studied item, ω_{ij}, have been removed. This model comparison is conducted separately for each studied item; again, a significant difference between models is indicative of the item exhibiting DIF.

At both stages of this analysis, χ^2 difference testing is required. As mentioned earlier, in M*plus*, the XWITH command is used to create the models' interaction terms. XWITH is used in conjunction with the "TYPE=RANDOM" command to estimate a random-effects model (Muthén, 2004), which "precludes the calculation of standardized coefficients and chi-square and related fit statistics" (Muthén, 2009, para. 1). To further complicate matters, as Satorra and Bentler (2010) explained,

> it frequently happens that two nested models, M_0 and M_1, are compared using estimation methods that are nonoptimal (asymptotically) given the distribution of the data; e.g., maximum likelihood (ML) estimation is used when the data are not multivariate normal. In those circumstances, the usual chi-square difference test

$T_d = T_0 - T_1$, based on the separate models' goodness of fit test statistics, is not χ^2 distributed. (p. 243)

Instead, the Satorra-Bentler χ^2 test of significance (Satorra & Bentler, 2001; see, also, Muthén & Muthén, n.d.) is used to compare models. This difference testing is conducted using the loglikelihoods – which follow a χ^2 distribution (Muthén, 2009) – and the scaling correction factors associated with each model. To conduct this χ^2 test, the difference test scaling correction, c_d, is calculated as

$$c_d = \frac{p_0 c_0 - p_1 c_1}{p_0 - p_1},$$

where p_0 is the number of parameters in the null (nested) model, p_1 is the number of parameters in the alternative (comparison) model, c_0 is the scaling correction factor of the null model, and c_1 is the scaling correction factor of the alternative model. This χ^2 test takes the form of

$$TR_d = \frac{-2\left(L_0 - L_1\right)}{c_d},$$

where TR_d is the Satorra-Bentler scaled χ^2 difference statistic, L_0 is the loglikelihood of the null model, and L_1 is the loglikelihood of the alternative model. The degrees of freedom for this χ^2 test are equal to $p_1 - p_0$ (UCLA Statistical Consulting Group, n.d.). It should be noted that, like other χ^2 tests of significance, the Satorra-Bentler χ^2 is sensitive to sample size (Pendergast et al., 2017).

MIMIC models have several advantages over other DIF detection methods. Unlike the MH and GMH methods, in which examinees are matched on summed scores, MIMIC models use latent-variable matching, a procedure which is likely to be more accurate (Woods et al., 2009). These models have been extended to include multiple groups, as well as multiple

grouping factors (e.g., age and ethnicity; Woods et al., 2009). Interaction terms (e.g., gender by ethnicity) may also be tested for significance. MIMIC models may include both continuous and categorical covariates to account for other sources of variation (Finch, 2005). [Conversely, researchers may use MIMIC models to control for DIF (Fleishman, Spector, & Altman, 2002).] Additionally, as only one set of item parameters are estimated for the entire sample – as opposed to one set per group – the sample size requirements are lower for MIMIC than for some other DIF detection methods. Furthermore, there is some evidence that MIMIC models exhibit more accurate uniform DIF detection for dichotomous items when focal group sample sizes are small (Woods, 2009b). MIMIC models have also been shown to have a well-controlled Type I error rate and high power (Shih & Wang, 2009).

MIMIC models also have a few disadvantages in regard to DIF detection. Perhaps the most significant is the fact that a researcher cannot test for DIF across all items simultaneously; such a model would not be identified; "at least one DIF-free [anchor] item is needed to define the factor on which the groups are matched" (Woods et al., 2009, p. 323). This requires the researcher to compare nested models by conducting the χ^2 test of significance *for each item.* A second disadvantage of MIMIC models is the need to arbitrarily scale the latent response variables and the common factor. There are two options for scaling the latent response variables: constraining y_i^* to 1.0 for all items or constraining the residuals ε_i to have unit variance. There are also two methods for scaling the common factor: choosing a reference indicator or standardizing the common factor. In other words, the researcher may either fix the threshold and factor loading of one item while freely estimating the mean and variance of the latent factor or fix the mean and variance of the latent factor while freely estimating the item's threshold and factor loading. This creates four crossed scaling options from which the researcher may choose.

For a thorough explanation of these options, readers are referred to an excellent overview and demonstration by Kamata and Bauer (2008). A third disadvantage of MIMIC is that a reference group must still be arbitrarily selected by the researcher when dummy-coding the grouping variables. For example, if a researcher wished to study items for potential bias across White, Asian, Black, and Hispanic examinees, there would be separate factor loadings for each pairwise comparison. (Readers should note that MIMIC results are conceptually equivalent to GMH results, in which a significant omnibus finding indicates that there was a significant difference between at least one pair of groups.) Finally, as noted earlier, an inflation of the Type I error rate may be observed (Klein & Moosbrugger, 2000; Woods & Grimm, 2011), because the model estimation method violates the assumption of normally-distributed variables when the model includes dummy-coded grouping variables.

Despite these drawbacks to MIMIC models, there is some evidence that they may be a powerful tool for DIF detection. Finch (2005), in studying the detection of uniform DIF, found that MIMIC models had comparable performance, as measured by power and Type I error, with the MH, the IRT likelihood ratio (Thissen, Steinberg, & Gerrard, 1986), and SIBTEST (Shealy & Stout, 1993) methods for 50-item exams and when there is no pseudo-guessing parameter. Finch (2005) also found that the MIMIC model was less sensitive to anchor item contamination and small focal group sample sizes. Shih and Wang (2009) found that, when using MIMIC models iteratively to find a pure anchor and then assessing items for uniform DIF, a four-item anchor resulted in high power with equal sample sizes of 1,000 and acceptable Type I error rates across all conditions, even when 40% of the test items exhibited DIF. Chun et al. (2016) found that, when factor variance across groups was equal, using the sequential-free baseline approach to (non)uniform DIF detection, in which the most discriminating DIF-free item is used as the

anchor for subsequent DIF detection tests, an inappropriate anchor was chosen only 1% of the time. Furthermore, Type I error rates were acceptable, and power was high when detecting uniform DIF in groups of $N = 250$. It bears mention that all these authors have noted the dearth of simulation studies for the MIMIC method and have recommended continued study across varying conditions.

Base/Omnicultural Reference Group

One distinct difference between the (MF)MG-NCDIF, GMH, and MIMIC methods is the composition of the reference group. GMH and MIMIC have historically employed a traditional reference group. However, as noted earlier, (MF)MG-NCDIF uses a base group. Other authors have discussed the use of a base group, although sometimes under a different heading. The term *composite*, for example, has also been used (e.g., Sari & Huggins, 2015). Ellis and Kimmel (1992) perhaps described this group in the most informative way when they deemed it an *omnicultural* reference group. As they explained, traditional pairwise DIF analysis indicates only when the focal group's response pattern differs from the reference group's pattern; these analyses fail to identify "a unique response pattern of one group to the exclusion of others" (p. 177). The ideal comparison, they argued, would be that of a focal group against an acultural reference, "a reference that is outside of and, therefore, uninfluenced by the culture itself" (p. 178). Although this acultural reference is an impossibility, the development and use of an omnicultural reference, which includes as many cultural-linguistic contributors as possible, is a close approximation of the acultural ideal: "the broader the composite [i.e., the omnicultural reference], the freer it is from any single culture's influence and the more decentered it is as a frame of reference" (p. 178). As base, composite, and omnicultural reference groups are synonymous, these terms will be used interchangeably.

As with any statistical procedure, there are advantages and disadvantages to the use of a base group in DIF analysis. Sari and Huggins (2015) identified six critical advantages with a base-group analysis. One advantage is that focal groups may simultaneously be defined by multiple factors; for example, the response patterns of Hispanic females may be compared with the base. A second advantage is that the need to arbitrarily assign a reference group becomes obsolete. Third, unlike traditional pairwise comparisons, the composite approach makes use of the operational item parameters. Fourth, each DIF estimate in a base group analysis is specific to one group, not one group as compared with another group. A traditional DIF analysis may result, for example, in three estimates for Hispanic examinees (e.g., DIF for a Hispanic/White comparison, a Hispanic/Black comparison, and a Hispanic/Asian comparison), necessitating a search for Hispanic DIF patterns to determine whether there is a consistent (dis)advantage for Hispanic examinees. Composite results are easier to interpret than multiple pairwise results by providing a single measure of DIF for a focal group, "mak[ing] the group and direction of advantage very clear" (p. 672). As the number of groups increases, this advantage becomes more pronounced. Fifth, Type I error rates are lower for base group analyses than traditional pairwise analyses because fewer comparisons are made (except when pairwise comparisons with only one reference group are conducted) and sample sizes are larger; false discovery rates are also lower, for the same reasons. Finally, if more than two groups are examined, the sample size is larger for the composite than for a single reference group; therefore, when holding the effect size constant, power is higher for composite analyses.

Sari and Huggins (2015) also highlighted three disadvantages to the use of a base group in DIF analyses. As opposed to traditional pairwise comparisons in which independent observations are maintained, the use of a base group introduces dependence when examinees

from the focal group are also included in the omnicultural base group. [Ellis and Kimmel (1992) also recognized this potential drawback and recommended excluding the focal group from the base while still including at least five cultural-linguistic groups in the base, if possible.] Second, there is a wider variety of DIF detection methods for pairwise analyses (e.g., Mantel-Haenszel and Lord's chi-square test). These methods have been more thoroughly researched and their efficacy tested using a traditional reference group. Third, if the DIF index is unweighted, then the composite is highly sensitive to the sample sizes of groups (e.g., the ICC of examinees without disabilities will rarely show DIF when the composite approach is used because the examinees in this group account for most of the composite). This may be a problem as composite groups are often used in score equity assessment.

Additionally, Sari and Huggins (2015) explained that the change from a traditional reference group to a base group is accompanied by a corresponding change in the definitions of "fairness" and "bias." Whereas traditional pairwise analysis compares the ICCs of two groups, the composite approach compares a focal group ICC with the ICC of all examinees (i.e., the operational ICC). DIF analysis is used to identify fairness as a lack of bias. *Fairness* in pairwise analyses means that the ICC for one group is equivalent to the ICC for each of the other groups; *fairness* in composite analyses means that the ICC for one group is equivalent to the operational ICC. Pairwise results indicate where "group-to-group invariance" is tenable; composite results indicate where the assumption of "group item parameters being invariant to the operational item parameters" (p. 671) – which are used for estimating scores – is tenable.

Similarly, Mayhew and Simonoff (2015) argued that the type of coding used – effect coding versus reference coding – affects interpretations of race-based analyses. In summarizing Mayhew and Simonoff's work, Rios-Aguilar (2014) stated

> they propose the use of *effect coding* rather than *reference coding* when comparing the outcomes of different racial/ethnic groups. Effect coding allows the comparison in outcomes of interest of groups to the grand mean, whereas reference coding compares groups of students with each other, thus allowing scholars to better understand the outcomes of certain groups of students compared to an institutional average expected for all students, rather than simply comparing underrepresented students to the outcomes of White students. (pp. 98-99)

Therefore, Mayhew and Simonoff advocated for the use of effect coding when studying race:

> By removing the idea of a reference group and by interpreting categorical effects as those that differ from an overall level, effect coding may equip quantitative criticalists . . . with the language they need to start making more informed choices regarding the use of statistics in understanding race and its effects on a variety of outcomes of interest. (p. 174)

It should be noted that Mayhew and Simonoff acknowledge the quantitative validity of using reference coding; however, they ask the question "From an inclusive perspective, what are the implications of consistently essentializing the voices of any group of students as the benchmark for understanding racial differences?" (p. 174).

Of course, pairwise and base-group methods may or may not identify the same set of items as exhibiting DIF for the same groups. The intention behind sharing these perspectives is not as an argument against the traditional pairwise approach – indeed, the two approaches can complement each other – but as a recommendation to choose the approach based on (a) the definition of fairness that is applicable to a study and (b) the consideration of the relative

(dis)advantages of each approach. This is an echo of the recommendations of Ellis and Kimmel (1992), Huggins and Penfield (2012), and Sari and Huggins (2015).

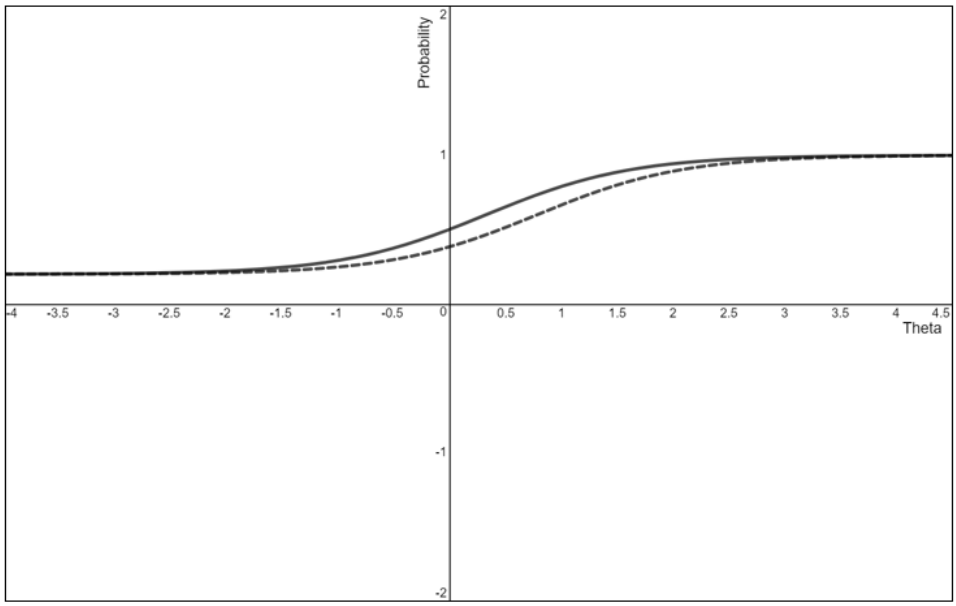

Figure 1. Uniform DIF. The solid line has parameters of a = 0.90, b = 0.50, and c = 0.20. The dashed line has parameters of a = 0.90, b = 1.20, and c = 0.20.

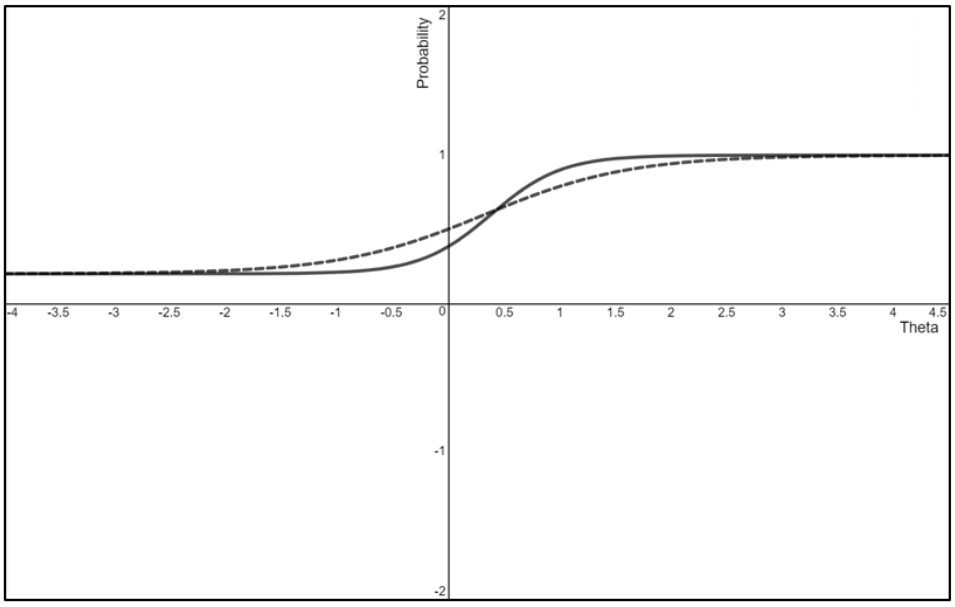

Figure 2. Non-Uniform DIF. The solid line has parameters of a = 1.85, b = 0.50, and c = 0.20. The dashed line has parameters of a = 0.90, b = 1.20, and c = 0.20.

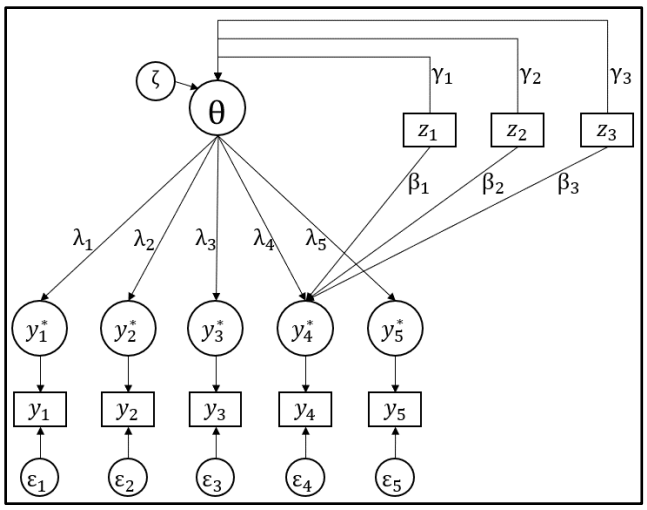

Figure 3. Five-Item Uniform DIF MIMIC Model with Three Dummy-Coded Grouping Variables. Item 4 is the studied item. A significant β_i value indicates the presence of uniform DIF for the z_i/reference group comparison.

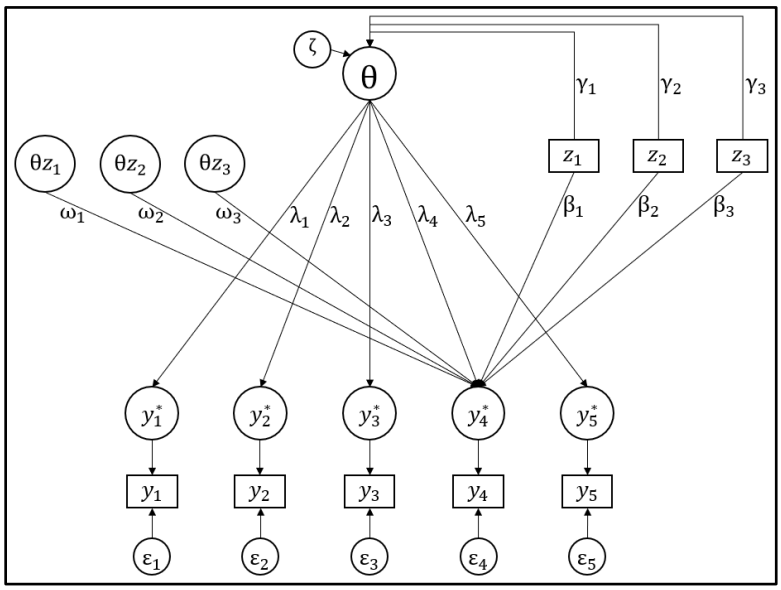

Figure 4. Five-Item Non-Uniform DIF MIMIC Model with Three Dummy-Coded Grouping Variables. Item 4 is the studied item. A significant β_i value indicates the presence of non-uniform DIF for the z_i/reference group comparison.

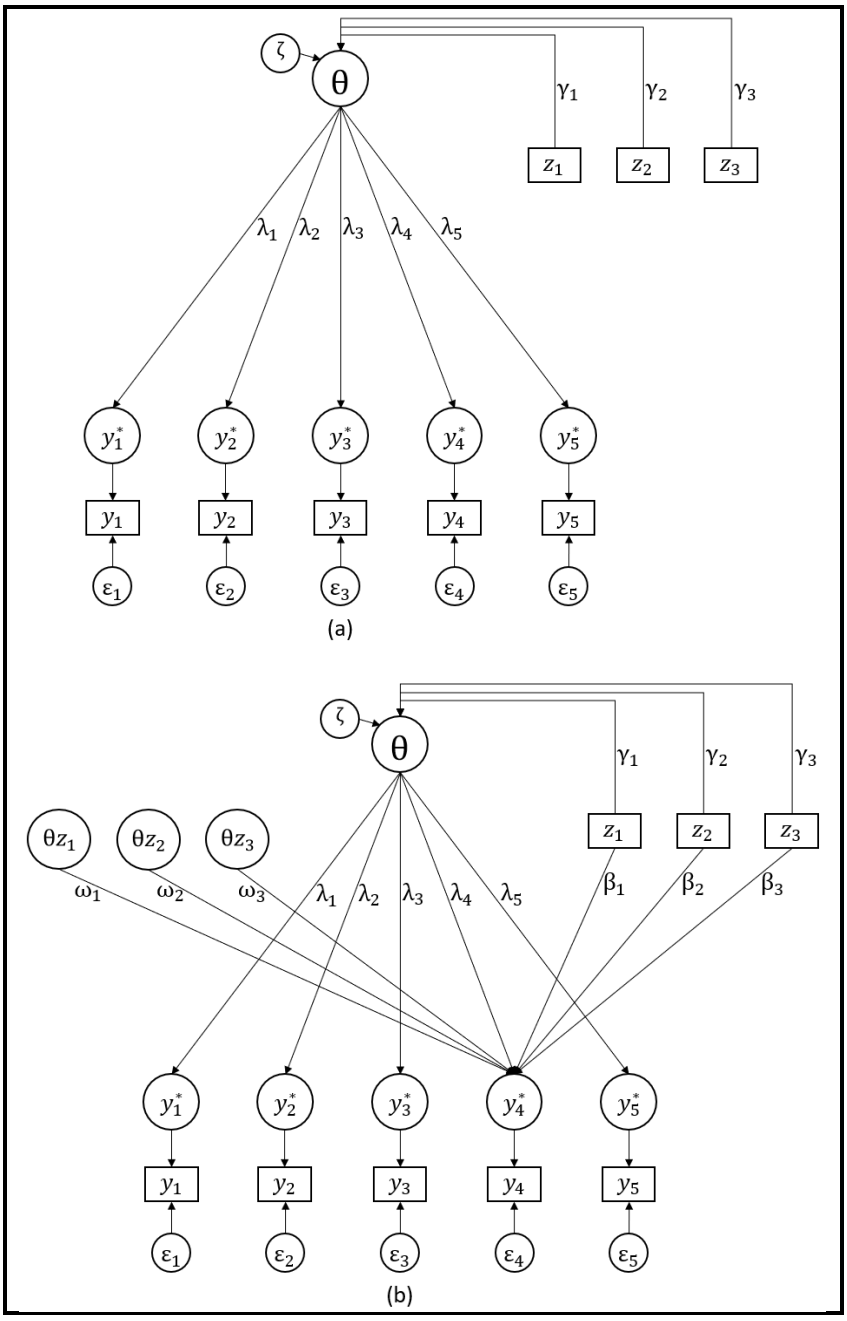

Figure 5. Constrained Baseline Approach for Stage 1 DIF Testing. Item 4 is the studied item.

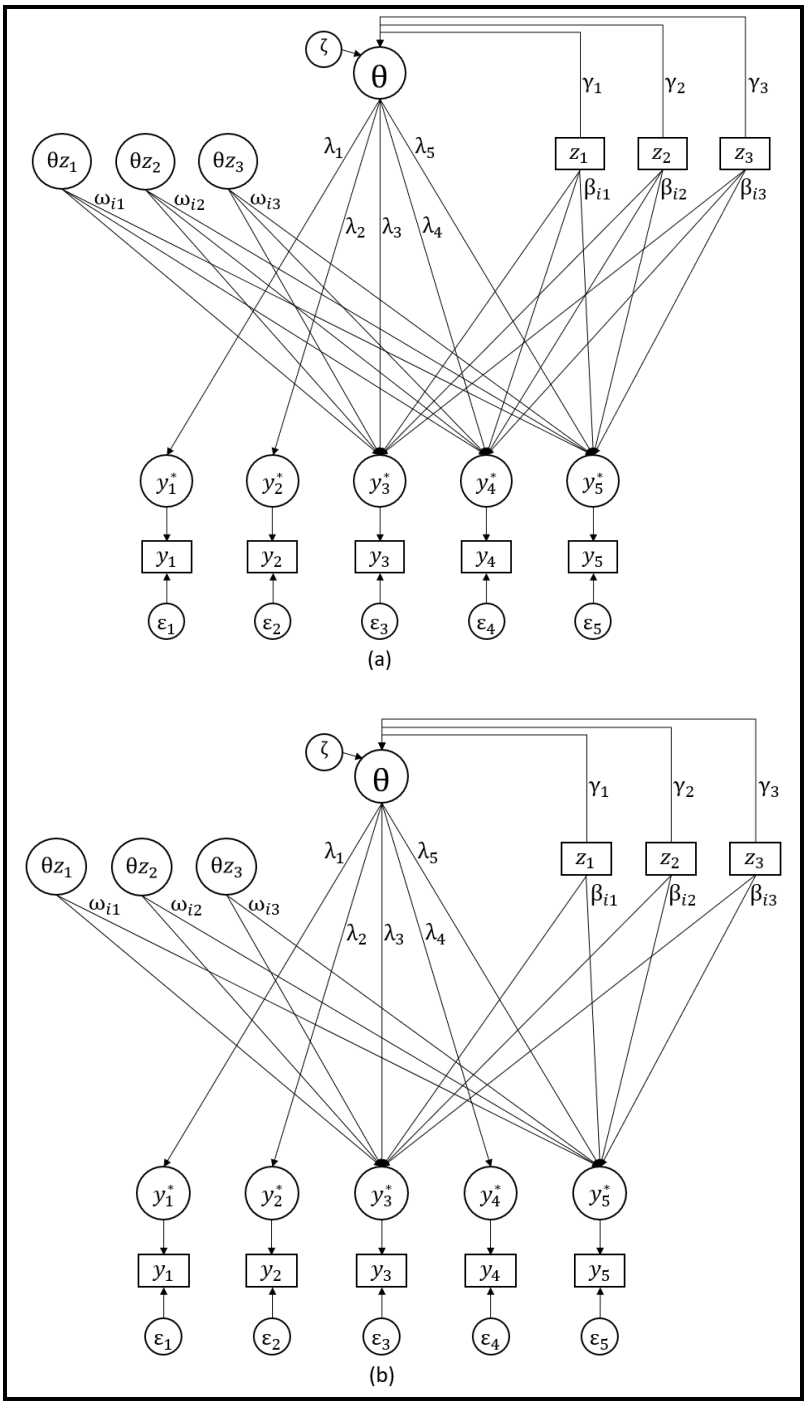

Figure 6. Free Baseline Approach for Stage 2 DIF Testing. Item 4 is the studied item; Items 1 and 2 are the anchors.

3 METHODOLOGY

Oshima et al. (2015) conducted a simulation study to assess the performance of the MG-NCDIF index; however, the performance of the MG-NCDIF index was not compared to other indices. To answer the research questions identified for this study, Oshima et al.'s study conditions were replicated. The simulated test consisted of 30 dichotomous items, of which zero (0%) or six (20%) contained DIF, depending on the condition. DIF was embedded in the same manner as Seybert and Stark (2012), based on their *a* and *b* item parameters. Tables 1 and 2 indicate which items were embedded with DIF and present the item parameters. In accordance with a critical race theory/QuantCrit framework, all analyses were conducted using a base reference group. The following studied conditions were manipulated.

Number of Groups

The analysis was comprised of two levels of number of groups: three groups and five groups.

Sample Size

Two sample size conditions were included in this study: equal (N = 1000) and unequal (N = 500 or N = 1000). For unequal sample sizes, Groups 2 and 3 each had N = 500 and all remaining groups had N = 1000 for both the three-group and five-group conditions. The base group sample size for each condition was set as the average of the sample sizes of the other groups in the condition. Therefore, for equal-size conditions, N = 1000 for the base group; for unequal-size conditions, N = 667 and N = 800 for the three-group and five-group conditions, respectively.

Type of DIF

Both uniform and non-uniform DIF were analyzed. Uniform DIF was simulated by increasing or decreasing the *b* parameter. Non-uniform DIF was simulated by increasing or decreasing (a) the *a* parameter only or (b) both the *a* and *b* parameters.

DIF Patterns

For uniform DIF, the *b* parameter was calculated by adding or subtracting 0.7 to/from the original parameter value. In the three-group condition, three patterns were used: (0/0/0), (0/+0.7/0), and (0/+0.7/−0.7). The first pattern (0/0/0) indicates the absence of DIF, as zero was added to the original *b* parameter for Group 1, zero was added to the *b* parameter for Group 2, and zero was added to the *b* parameter for Group 3. The second pattern (0/+0.7/0) indicates that 0.7 was added to the *b* parameter for Group 2, making the item more difficult for that group. The third pattern (0/+0.7/−0.7) indicates that 0.7 was added to the *b* parameter for Group 2, making the item more difficult for that group, while 0.7 was subtracted from the *b* parameter for Group 3, making the item easier for that group. In the five-group condition, the following uniform DIF patterns were applied: (0/+0.7/0/0/0), (0/+0.7/+0.7/0/0), and (0/+0.7/+0.7/−0.7/−0.7).

For non-uniform DIF via manipulation of the *a* parameter only, 0.4 was added to or subtracted from the original parameter value. In the three-group condition, the patterns were (0/0/0), (0/−0.4/0), and (0/−0.4/+0.4). The first pattern indicates that the item was equally discriminating for all three groups; the second pattern indicates that the item was less discriminating for Group 2; the third pattern indicates that the item was least discriminating for Group 2 but most discriminating for Group 3. In the five-group condition, the patterns were (0/−0.4/0/0/0), (0/−0.4/−0.4/0/0), and (0/−0.4/−0.4/+0.4/+0.4).

For non-uniform DIF via manipulation of both the *a* and *b* parameters, the patterns above were combined, respectively. For example, in the second pattern of the three-group condition, the *b* parameter was adjusted using (0/+0.7/0) and the *a* parameter was simultaneously adjusted using (0/−0.4/0).

These patterns may be classified as unidirectional or bidirectional. Unidirectional patterns were defined as the patterns where, when the *a* and/or *b* parameters were manipulated for one or more groups, these manipulations shared the same sign (+ or −). Cases in which only one group's parameters were manipulated were also considered unidirectional. Bidirectional patterns were defined as the patterns where, when the *a* and/or *b* parameters were manipulated for one or more groups, these manipulations had differing signs. For example, (0/−0.4/−0.4/0/0) is a unidirectional pattern and (0/−0.4/−0.4/+0.4/+0.4) is a bidirectional pattern.

Impact

The simulation included three levels of impact: no impact, unidirectional impact, and bidirectional impact. For the no-impact condition, each group had an ability distribution of $N(0, 1)$. For unidirectional impact, all groups had ability distributions of $N(0, 1)$ except for Groups 2 and 3, which followed distributions of $N(-0.5, 1)$. In other words, Groups 2 and 3 had lower ability than the remaining groups. For bidirectional impact, all groups had ability distributions of $N(0, 1)$ except for Group 2, which followed a distribution of $N(-0.5, 1)$, and Group 3, which followed a distribution of $N(+0.5, 1)$. In this case, Group 3 had the highest mean ability level of all the groups; Group 2 had the lowest mean ability of all the groups. Oshima et al. (2015) found negligible difference between unidirectional and bidirectional impact in their three-group design, so they elected not to simulate bidirectional impact in the five-group condition. For consistency,

the bidirectional impact condition was also omitted from the five-group design in the current study.

Data Simulation

Datasets were simulated for each crossed condition using R 4.0.3 (R Core Team, 2020) within the RStudio 1.3.1093 (RStudio Team, 2020) integrated development environment. Initial item parameters were maintained or manipulated, depending on the DIF pattern condition. Theta values were randomly generated from the random normal distribution for each examinee based on the groups' ability means and standard deviations, as discussed previously. The resulting item parameters and theta values were applied to a three-parameter logistic (3PL) model with a scaling constant of $D = 1.7$ to calculate the probabilities of a correct response for each examinee for each item in the crossed condition. Random numbers – one per examinee per item – were then generated from a uniform distribution [$X \sim U(0, 1)$]. Each probability was compared against the corresponding random uniform value. If the probability was greater, the examinee was assigned a correct response (i.e., a "1") for that item; if not, the examinee was assigned an incorrect response (i.e., a "0") for that item. Each base group was comprised of a random sample of examinees and their responses from the crossed condition.

MG-NCDIF

MG-NCDIF was not re-simulated in the current study, as all study conditions were identical to those of Oshima et al. (2015). However, as readers may be interested in a summary of their methodology, particularly as it compares to the current study, a brief overview is included here. In Stage 1, BILOG-MG3 (Zimowski, Muraki, Mislevy, & Bock, 2003) was used to calibrate simulated item responses, item parameters were then estimated and placed on a common scale (that of the base group), and a modified version of the program "DIFCUT"

(Nanda, Oshima, & Gagne, 2005) was run in SAS (SAS Institute Inc., 2012) to test for significant MG-NCDIF at $\alpha = .05$. Items found to be DIF-free were then used as anchors in Stage 2, in which the TCC linking procedure (Stocking & Lord, 1983) was employed to retrieve purified linking coefficients. Next, DIFCUT analyses were repeated with the purified linking coefficients, identifying a final set of items exhibiting significant MG-NCDIF. In both stages, the IPR method (Oshima et al., 2006) was used for significance testing. As Oshima et al. found 600 pairs of simulated item parameters to be sufficient, DIFCUT was set to simulate 600 pairs of item parameters using the variances, covariances, and theta estimates of the base group. Readers interested in a full treatment of the methodology are, of course, referred to the source article.

It bears mention that Oshima et al. (2015) conducted these MG-NCDIF analyses using (a) significance tests only and (b) significance tests in conjunction with the effect size measure developed by Wright and Oshima (2015). However, to provide consistency with the GMH and MIMIC methods, which did not make use of an effect size, the current study focuses on the significance-test-only MG-NCDIF results.

GMH

For the GMH index, R 4.0.3 (R Core Team, 2020) within the RStudio 1.3.1093 (RStudio Team, 2020) integrated development environment was used to detect DIF. Specifically, the "difGMH" function of the "difR" package (Magis et al., 2010) was employed to conduct a two-stage analysis, which has been shown to produce a slight improvement over a one-stage GMH analysis (Wang & Su, 2004). At Stage 1, all items were tested for the purpose of identifying DIF-free items, which became the anchor items for Stage 2. Stage 2 consisted of two parts: *partial purification* and *full purification*. To achieve partial purification, each item that exhibited DIF in Stage 1 was retested using the anchor items; to achieve full purification, each anchor item was

then retested against the remaining anchor items (Fikis & Oshima, 2017). Per the recommendation of Zwick, Donoghue, and Grima (1993), the studied item was always included in the total (matching) score. Procedures were embedded in the GMH code to address two special circumstances in this two-stage process: failure to recover any anchor items or recovery of a single anchor item. If no anchors were identified in Stage 1, which rendered further testing unnecessary, the Stage 1 results were treated as the final results *for the dataset*. If a single anchor item was identified, it was not possible to re-test that item in Stage 2 because there were no remaining anchor items; in this instance, the Stage 1 results were treated as the final results *for that item*.

Kim and Oshima (2012) recommended the use of a multiple testing adjustment method when conducting DIF studies, given that significance is tested for each item. In his discussion of multiple-group DIF analyses, Penfield (2001) made a similar recommendation, given that significance tests were required for each pairwise comparison. In their simulation study, Kim and Oshima (2012) found that the Benjamini-Hochberg (BH) false discovery rate procedure (Benjamini & Hochberg, 1995) provided a sufficient balance of Type I error rate control and power when using the MH method (Mantel & Haenszel, 1959) to detect DIF. Additionally, the BH procedure has been used in several MIMIC studies (see below). Therefore, at both stages of the GMH analysis, the observed *p* values were adjusted using the BH procedure at $\alpha = .05$.

All “difGMH” default settings were applied to these analyses. It bears acknowledgement that the remaining default settings may not be comparable to the settings used for the MG-NCDIF and MIMIC methods. However, assessment companies and researchers are not likely to set the GMH analysis options to match MG-NCDIF or MIMIC options (or options of other DIF

indices). Thus, the results of this study reflect what analysts would see if they used GMH as an isolated measure in practice with operational data.

MIMIC

For the MIMIC index, M*plus* 8.4 (Muthén & Muthén, 2007) was used to identify a purified set of anchors and detect DIF. To automate these analyses, R 4.0.3 within the RStudio 1.3.1093 integrated development environment was used. Specifically, the "MplusAutomation" package (Hallquist & Wiley, 2018) was employed to write the M*plus* input code, call M*plus* to run the analyses, and read in the M*plus* output files.

As mentioned earlier, in M*plus*, the XWITH command is used to create the models' interaction terms. XWITH was used in conjunction with the "TYPE=RANDOM" command to estimate a random-effects model (Muthén, 2004). Analyses conducted with XWITH utilized the LMS algorithm and robust maximum likelihood (MLR) estimation with numerical integration (Muthén & Muthén, 2017). To ensure that the model was identified, the mean and variance of the latent factor, y_i^*, were constrained to zero and one, respectively. This freed the Item 1 factor loading, permitting the item to be studied for DIF. All other M*plus* default settings were applied.

Per the recommendations of Lopez Rivas et al. (2009) and Chun et al. (2016), two-stage omnibus testing for uniform and/or non-uniform DIF was conducted. At Stage 1, a constrained baseline approach was used to identify anchor items (see Figure 5); at Stage 2, a free baseline approach was used to identify items functioning differentially (see Figure 6).

Researchers have examined power and Type I error rates for various DIF detection methods when the number of anchor items has been limited in an effort to reduce the risk of anchor contamination, finding that one to five anchors yielded desirable results (e.g., Lopez Rivas et al., 2009; Meade & Wright, 2012). Since then, researchers have studied the performance

of a short anchor with the MIMIC method. Designating four items as anchors based on the first stage of testing, Rebouças and Cheng (2019) found that highly-discriminating anchors were more likely to flag DIF items, while power was reduced when anchors had low discrimination. Shih and Wang (2009) found that, in MIMIC analyses, anchor sets of one, two, four, and ten items were sufficient to detect DIF items while exhibiting well-controlled Type I error rates; a four-item anchor set yielded high power rates that were similar to the ten-item anchor set power rates. Thus, they recommended a four-item anchor set, which was 10-20% of their test length, depending on the condition. Other MIMIC studies have made use of a short anchor while examining other conditions of interest. For example, Chun et al. (2016) used a single highly-discriminating item for the anchor when investigating MIMIC DIF detection performance with (non)uniform DIF in a multiple-group context. Wang and Shih (2010) used four anchors when comparing various MIMIC methods to detect DIF in polytomous items. Woods and Grimm (2011) used one-third of simulated assessment items for anchoring when they introduced the MIMIC model for non-uniform DIF detection and a three-item anchor in their empirical example. Based on this clear precedent for limiting the number of anchors, as many as five DIF-free items were selected as anchors in the current study. In cases where more than five items were found to be DIF-free, the five with the highest discrimination (loading) values were assigned as the anchors. At least one item must be assigned as an anchor to achieve model identification in Stage 2; therefore, if no anchor was recovered in the first stage, Item 5, the most discriminating DIF-free item, was designated as the anchor.

At both stages of analysis, the Satorra-Bentler χ^2 test of significance (Satorra & Bentler, 2001) was used to compare nested models. These model comparisons – in addition to data aggregation and calculation of outcome measures (see next section) – were conducted using R in

the RStudio environment. As discussed in the previous section, Kim and Oshima (2012) and Penfield (2001) recommended the use of a multiple testing adjustment method when conducting DIF studies. The BH procedure (Benjamini & Hochberg, 1995), which Kim and Oshima (2012) found to be effective when using the MH method (Mantel & Haenszel, 1959) to detect DIF, has also been used in several MIMIC studies to date, including Woods (2009a), Woods (2009b), Woods and Grimm (2011), and Woods et al. (2009). Therefore, at both stages of the MIMIC analysis, the observed *p* values were adjusted using the BH procedure at $\alpha = .05$.

Outcomes

The manipulation of these conditions created 82 crossed conditions, which were replicated 100 times each. For each crossed condition, Type I error rates and power were calculated as the outcome measures. Type I error was calculated as the mean of the number of times an item was incorrectly flagged as exhibiting DIF across the 100 replications. Power was calculated as the mean of the number of correct DIF identifications across the 100 replications.

Table 1
Item Parameters Used in Generating DIF Conditions – Three-Group Condition

	G1			G2		G3	
Item	*a*	*b*	*c*	*a*	*b*	*a*	*b*
1	0.49	−0.07	0.19				
2	0.92	0.21	0.15				
3	1.26	0.54	0.05				
4	0.61	−0.03	0.18	(−0.4)	(+0.7)	(+0.4)	(−0.7)
5	1.74	0.01	0.12				
6	0.50	1.96	0.12				
7	0.96	0.04	0.13	(−0.4)	(+0.7)	(+0.4)	(−0.7)
8	0.59	−0.09	0.18				
9	0.82	−1.16	0.17				
10	1.26	0.02	0.11				
11	0.82	0.20	0.07				
12	0.75	−0.43	0.15				
13	1.49	−0.06	0.09	(−0.4)	(+0.7)	(+0.4)	(−0.7)
14	0.97	−0.34	0.12				
15	1.49	0.05	0.12				
16	0.89	−0.25	0.15				
17	1.45	0.06	0.07				
18	0.75	0.31	0.18				
19	1.43	0.04	0.08	(−0.4)	(+0.7)	(+0.4)	(−0.7)
20	0.60	0.13	0.22				
21	0.83	0.52	0.09				
22	0.56	−0.96	0.19	(−0.4)	(+0.7)	(+0.4)	(−0.7)
23	0.67	−0.79	0.20				
24	0.70	0.37	0.18				
25	1.03	−0.71	0.14				
26	0.89	−0.19	0.21				
27	1.23	0.74	0.06				
28	0.90	−0.44	0.18	(−0.4)	(+0.7)	(+0.4)	(−0.7)
29	1.23	−0.17	0.12				
30	0.69	0.53	0.17				

Note. The bold numbers (4, 7, 13, 19, 22, and 28) indicate DIF items. The number in the parentheses under G2 and G3 indicates the value subtracted from or added to the item parameters for G1 to embed DIF, when applicable, on either the *a* parameter, the *b* parameter, or both. From "Multiple Group Noncompensatory Differential Item Functioning in Raju's Differential Functioning of Items and Tests," by T. C. Oshima, K. Wright, and N. White, 2015, *International Journal of Testing, 15*, pp. 254-273. Reprinted with permission.

Table 2
Item Parameters Used in Generating DIF Conditions – Five-Group Condition

	G2		G3		G4		G5	
Item	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*
1								
2								
3								
4	(−0.4)	(+0.7)	(−0.4)	(+0.7)	(+0.4)	(−0.7)	(+0.4)	(−0.7)
5								
6								
7	(−0.4)	(+0.7)	(−0.4)	(+0.7)	(+0.4)	(−0.7)	(+0.4)	(−0.7)
8								
9								
10								
11								
12								
13	(−0.4)	(+0.7)	(−0.4)	(+0.7)	(+0.4)	(−0.7)	(+0.4)	(−0.7)
14								
15								
16								
17								
18								
19	(−0.4)	(+0.7)	(−0.4)	(+0.7)	(+0.4)	(−0.7)	(+0.4)	(−0.7)
20								
21								
22	(−0.4)	(+0.7)	(−0.4)	(+0.7)	(+0.4)	(−0.7)	(+0.4)	(−0.7)
23								
24								
25								
26								
27								
28	(−0.4)	(+0.7)	(−0.4)	(+0.7)	(+0.4)	(−0.7)	(+0.4)	(−0.7)
29								
30								

Note. The item parameters for G1 are shown in Table 1. The value in the parentheses indicates the number subtracted from or added to the item parameters for G1, when applicable. From "Multiple Group Noncompensatory Differential Item Functioning in Raju's Differential Functioning of Items and Tests," by T. C. Oshima, K. Wright, and N. White, 2015, *International Journal of Testing, 15,* pp. 254-273. Reprinted with permission.

4 RESULTS

Power and Type I error rates were calculated for each of the 82 crossed conditions. The results for the GMH three-group and five-group conditions are shown in Tables 3 and 4, respectively. MIMIC results for the three-group and five-group analyses are provided in Tables 5 and 6, respectively. For ease of comparison, the MG-NCDIF results from Oshima et al.'s (2015) study have been cited, in Tables 7 and 8, with the lead author's permission. Although Oshima et al. provided power and Type I error results both with and without the use of an effect size measure, only the significance testing results are included herein, to provide consistency with the GMH and MIMIC results. It should be noted that, while Oshima et al. reported power and Type I error to the hundredths place for MG-NCDIF, power and Type I error for GMH and MIMIC are reported to the thousandths place for greater precision. This precision was particularly important for these two methods for conditions with lower Type I error rates (many of which fell below .010) and for higher MIMIC power rates (many of which exceeded .990).

Overall, GMH exhibited the least consistent power, which ranged from .025 to 1.000, with the lowest observed power rates being associated with the non-uniform DIF condition in which only the *a* parameter was manipulated, particularly in cases where group sizes were unequal and mean ability varied between groups. The same was found to be true for MG-NCDIF, albeit with a slightly more consistent range of power (.15 to 1.00). Although the MIMIC conditions with the lowest power followed the same pattern, this method exhibited more consistent power, with values ranging from .340 to 1.000 and with only five of the 82 crossed conditions falling below a power of .500. (Comparatively, power fell below .500 for 17 of the MG-NCDIF conditions and 24 of the GMH conditions.) While all three methods exhibited power of .900 or higher in particular cases, MG-NCDIF did so across only 11 of the crossed conditions,

as compared with GMH and MIMIC, which exhibited power of at least .900 across 35 and 50 of the crossed conditions, respectively.

When it came to Type I error rates, MG-NCDIF exhibited inflated error, exceeding the .05 nominal alpha level 44 times. In regard to Type I error rates for GMH and MIMIC, the results were mixed. On one hand, the GMH index exhibited better Type I error control in nearly two and a half times as many of the crossed conditions as the MIMIC index at the .05 level. Similarly, GMH Type I error rates fell below .01 in 58 of the crossed conditions, while MIMIC results showed a Type I error rate below .01 in 39 of the crossed conditions. On the other hand, the MIMIC method Type I error rates never exceeded the nominal alpha level, while GMH Type I error rates exceeded .05 three times, going as high as .089. This occurred in the five-group equal sample size condition where (a) there was no impact and the DIF pattern was (0/−0.4/−0.4/+0.4/+0.4) for the *a* parameter and (0/+0.7/+0.7/−0.7/−0.7) for the *b* parameter, (b) there was no impact and the DIF pattern was (0/+0.7/+0.7/−0.7/−0.7) for the *b* parameter, and (c) impact was present and the DIF pattern was (0/+0.7/+0.7/−0.7/−0.7) for the *b* parameter.

As explained earlier, all items were tested at Stage 1 for the purpose of identifying DIF-free items to serve as anchors at Stage 2, necessitating the need to develop procedures for special anchor-identification circumstances. For the GMH index, Stage 2 consisted of two parts: *partial purification* and *full purification*. To achieve partial purification, each item that exhibited DIF in Stage 1 was retested using the anchor items; to achieve full purification, each anchor item was then retested against the remaining anchor items (Fikis & Oshima, 2017). Procedures were embedded in the GMH code to address two special circumstances in this two-stage process: failure to recover any anchor items or recovery of a single anchor item. If no anchors were identified in Stage 1, which rendered further testing unnecessary, the Stage 1 results were treated

as the final results *for the dataset*. If a single anchor item was identified, it was not possible to re-test that item in Stage 2 because there were no remaining anchor items; in this instance, the Stage 1 results were treated as the final results *for that item*. Failure to recover any anchor items occurred with 15 of the 8,200 simulated datasets – approximately 0.18% of the analyses – in the following crossed conditions, all of which had five groups and equal sample sizes:

- no impact with the uniform DIF pattern (0/+0.7/+0.7/−0.7/−0.7) for the *b* parameter (six datasets),
- no impact with the non-uniform DIF pattern (0/−0.4/−0.4/+0.4/+0.4) for the *a* parameter and (0/+0.7/+0.7/−0.7/−0.7) for the *b* parameter (six datasets), and
- impact present with the uniform DIF pattern (0/+0.7/+0.7/−0.7/−0.7) for the *b* parameter (three datasets).

Recovery of a single anchor item occurred with 53 of the 8,200 simulated datasets – approximately 0.65% of the analyses – in the following crossed conditions, all of which had five groups and equal sample sizes:

- no impact with the uniform DIF pattern (0/+0.7/+0.7/−0.7/−0.7) for the *b* parameter (14 datasets),
- no impact with the non-uniform DIF pattern (0/−0.4/−0.4/+0.4/+0.4) for the *a* parameter and (0/+0.7/+0.7/−0.7/−0.7) for the *b* parameter (22 datasets),
- impact present with the uniform DIF pattern (0/+0.7/+0.7/−0.7/−0.7) for the *b* parameter (13 datasets), and
- impact present with the non-uniform DIF pattern (0/−0.4/−0.4/+0.4/+0.4) for the *a* parameter and (0/+0.7/+0.7/−0.7/−0.7) for the *b* parameter (four datasets).

This means that, out of the 180,709 anchor items that were identified throughout this study, 53 (or 0.03%) of these items could not be re-tested for DIF in Stage 2.

For the MIMIC index, at least one item must have been assigned as an anchor to achieve model identification in Stage 2; therefore, if no anchor was recovered in the first stage, Item 5, the most discriminating DIF-free item, was designated as the anchor. This occurred with two of the 8,200 simulated datasets – approximately 0.02% of the analyses – in the five-group, equal sample size condition in which DIF was embedded in the *b* parameter (0/+0.7/+0.7/−0.7/−0.7): once when impact was present and once when it was not. Given the relatively small frequency with which these special anchor circumstances occurred, this topic will not be addressed in relation to the remaining results, which are organized by studied condition.

Number of Groups

The performance of the GMH and MIMIC DIF detection methods was affected by the number of groups. For both indices, the power tended to be the same or slightly higher in the five-group condition. For example, with the GMH index, in the non-uniform condition with no impact, unequal sample sizes, and a pattern of (0/−0.4/0) for the three-group condition and (0/−0.4/0/0/0) for the five-group condition, power equaled .120 and .132, respectively; the power of the MIMIC index in the same conditions was .382 and .395, respectively. Unfortunately, the GMH and MIMIC Type I error rates also tended to be the same or slightly higher in the five-group condition. These power and Type I error results were to be expected, given that GMH and MIMIC are both omnibus tests that at least one pairwise comparison will be found significant; the more comparisons that are made, the more likely an item is to be flagged – correctly or erroneously – for DIF. However, despite some higher flagging rates, both the three- and five-

group conditions followed the same performance patterns with respect to the other manipulated factors discussed below.

Oshima et al. (2015) found that the MG-NCDIF index was not sensitive to the number of groups studied. As they explained, care must be taken when interpreting the results, as the three-group condition and the five-group condition are not perfectly comparable due to the fact that the MG-NCDIF index is calculated by averaging the d^2 values of the pairwise comparisons. Take, for example, the case of unequal groups where no impact is present and uniform DIF has been embedded using the pattern (0/+0.7/0) for the three-group condition and the patterns (0/+0.7/0/0/0) and (0/+0.7/+0.7/0/0) for the five-group condition. It is expected for power to be lowest in the (0/+0.7/0/0/0) condition, with an increase in the (0/+0.7/0) condition, and the highest power in the (0/+0.7/+0.7/0/0) condition; indeed, this was the case, with observed power equaling .60, .63, and .71, respectively. The Type I error rates between these three example conditions were nearly identical: .03, .03, and .04, respectively.

Sample Size

For all three DIF detection methods, power and Type I error rates were generally higher for equal sample size conditions than the unequal sample sizes. (The differences in Type I error were especially pronounced for the MG-NCDIF index.) As noted earlier, all three methods are sensitive to sample size. The larger the sample size, the more likely an item is to be flagged – correctly or erroneously – for DIF. In the case of three-group analyses, the total sample size is 4,000 in the equal-groups conditions and 2,667 in the unequal-groups conditions (including the base group); in the case of five-group analyses, the total *N* counts are 6,000 and 4,800, respectively. These differences are substantial, and likely account for the observed differences.

With respect to the other manipulated conditions, datasets with equal and unequal sample sizes exhibited commensurate performance.

Type of DIF

Both uniform and non-uniform DIF were analyzed. Caution must be exercised when comparing the results for different types of DIF because the magnitude of DIF is greater in conditions where both the *a* and *b* parameters are manipulated to produce non-uniform DIF than conditions where a single parameter is manipulated to produce uniform DIF (via adjustment of the *b* parameter) or non-uniform DIF (via adjustment of the *a* parameter only). In other words, the effect of the type of DIF becomes confounded with the effect of the magnitude of DIF. Therefore, it was expected that the MG-NCDIF and MIMIC indices would exhibit the highest power when non-uniform DIF is modeled via manipulation of both the *a* and *b* parameters, with power decreasing for uniform DIF detection; the lowest power rates were expected to be associated with non-uniform DIF modeled by the *a* parameter only. For the GMH index, which is not designed to take differences in the *a* parameter into account, the magnitude of DIF did not increase when both parameters were manipulated; therefore, power for uniform DIF detection was expected to be highest.

As predicted, performance differences were observed between the types of DIF detected by the MIMIC index. MIMIC performed at its poorest when used to detect non-uniform DIF of the *a* parameter, with power ranging from .610 to .970 for equal groups and from .340 to .887 for unequal groups. The MIMIC method produced excellent results for uniform DIF detection, with power ranging from .995 to 1.000 for equal groups and from .937 to 1.000 for unequal groups. MIMIC results were also impressive when non-uniform DIF modeled by adjustments to both the

a and *b* parameters was simulated; power ranged from .997 to 1.000 for equal groups and from .948 to 1.000 for unequal groups.

Similarly, performance differences were observed between the types of DIF detected by the MG-NCDIF index. As was the case with the MIMIC method, MG-NCDIF performed at its poorest when used to detect non-uniform DIF of the *a* parameter; its power ranged from .27 to .63 for equal groups and from .15 to .55 for unequal groups. Uniform DIF detection with the MG-NCDIF method demonstrated lower power than the other two methods, with rates of .68 to .87 for equal groups and .60 to .88 for unequal groups. MG-NCDIF results improved when both the *a* and *b* parameters were manipulated to produce non-uniform DIF; under this condition, power ranged from .69 to 1.00 for equal groups and from .63 to 1.00 for unequal groups.

The GMH method produced excellent results for uniform DIF detection, with power ranging from .997 to 1.000 for equal groups and from .950 to 1.000 for unequal groups. However, as expected, GMH performed poorly when used to detect non-uniform DIF via manipulation of the *a* parameter only, with power ranging from .068 to .262 for equal groups and from .025 to .240 for unequal groups. When non-uniform DIF via manipulation of both the *a* and *b* parameters was present, the power of GMH improved, ranging from .793 to 1.000 for equal groups and from .647 to 1.000 for unequal groups.

DIF Patterns

For all three indices, within the unidirectional pattern, as the number of groups with DIF increased from one to two, the power increased as well. For example, in the GMH no-impact, unequal-groups condition with uniform DIF only, the single-group pattern of (0/+0.7/0/0/0) exhibited power of .980, while the multiple-group pattern of (0/+0.7/+0.7/0/0) demonstrated

power of 1.000. In other words, as the number of groups having DIF embedded in their parameters increased, DIF was more easily detected.

Similarly, for all three methods, the bidirectional patterns demonstrated higher power as compared to their unidirectional counterparts. For example, in the MIMIC no-impact, equal-groups condition with non-uniform DIF of the *a* parameter only, the bidirectional pattern of (0/−0.4/−0.4/+0.4/+0.4) exhibited higher power (.952) than the unidirectional pattern of (0/−0.4/−0.4/0/0), where power equaled .745. This is to be expected, given that in a bidirectional pattern, the difference in parameter values between the group with the highest parameter value and the group with lowest corresponding parameter value is twice that of the unidirectional pattern (0.8 versus 0.4 in this example). In other words, bidirectional DIF was more easily detected than unidirectional DIF for all studied methods.

It is interesting to note that the pattern of DIF also affected the Type I error rates for the GMH index. For GMH, the difference between the Type I error rate for an equal-groups condition and its unequal-groups counterpart was .004 or less approximately 78% of the time. For the remaining conditions, this difference ranged from .013 to .071, and all of these larger Type I error differences between an equal-groups condition and its unequal-groups counterpart occurred when the DIF pattern was either bidirectional uniform (via manipulation of the *b* parameter) or bidirectional non-uniform (via manipulation of both the *a* and *b* parameters). Furthermore, as stated earlier, the GMH Type I error rate did exceed the nominal alpha level of .05 for three of the 82 crossed conditions; all three of these cases occurred in conditions in which the DIF pattern was bidirectional. Thus, it seems that the GMH index is more sensitive to various DIF patterns than the other studied methods.

Impact

For the MIMIC method, the differences between impact conditions, when all other conditions were held constant, were small, a finding that may be attributed to the fact that group differences in mean ability are explicitly modeled by the paths from the grouping variables to the latent factor. Similarly, Oshima et al. (2015) concluded that the impact condition did not affect the performance of the MG-NCDIF index, owing to the efficacy of the two-stage linking procedure. Indeed, for MG-NCDIF the difference between impact conditions, when everything else was equal, was remarkably small.

However, the same does not appear to be true of the GMH statistic. In the GMH three-group conditions, power was typically highest when no impact was present and decreased when bidirectional impact was present, with unidirectional impact exhibiting the lowest power. This was consistent with the finding that, in the GMH five-group conditions, power was typically higher in the impact-free condition than the unidirectional-impact condition. This suggests that the flagging rates for items tended to increase when any type of impact was present but, as the GMH index does not partial out group mean differences from DIF, the resulting power was lowered, and Type I error increased.

Table 3
GMH Power and Type I Error Rates – Three-Group Condition

DIF Type	%DIF	Pattern	Impact	Equal *N*s (1000/1000/1000) P	Equal *N*s (1000/1000/1000) TIE	Unequal *N*s (1000/500/500) P	Unequal *N*s (1000/500/500) TIE
–	0	(0/0/0)	0/0/0	–	< .001	–	.000
			0/−0.5/−0.5	–	.001	–	.000
			0/−0.5/0.5	–	.001	–	.000
a	20	(0/−0.4/0)	0/0/0	.177	.002	.120	.002
			0/−0.5/−0.5	.068	.001	.025	.001
			0/−0.5/0.5	.103	.003	.040	.001
		(0/−0.4/0.4)	0/0/0	.212	.002	.178	.004
			0/−0.5/−0.5	.240	.005	.110	.004
			0/−0.5/0.5	.222	.005	.152	.003
b	20	(0/0.7/0)	0/0/0	1.000	.008	.980	.008
			0/−0.5/−0.5	.997	.006	.953	.008
			0/−0.5/0.5	1.000	.009	.950	.006
		(0/0.7/−0.7)	0/0/0	1.000	.016	1.000	.012
			0/−0.5/−0.5	1.000	.038	1.000	.008
			0/−0.5/0.5	1.000	.026	1.000	.009
a/*b*	20	(0/−0.4/0)	0/0/0	.890	.007	.843	.005
		(0/0.7/0)	0/−0.5/−0.5	.793	.006	.648	.008
			0/−0.5/0.5	.823	.009	.707	.005
		(0/−0.4/0.4)	0/0/0	1.000	.021	1.000	.008
		(0/0.7/−0.7)	0/−0.5/−0.5	1.000	.027	.998	.008
			0/−0.5/0.5	1.000	.035	1.000	.011

Note. DIF Type = manipulation of *a* parameter only, *b* parameter only, or both *a* and *b* parameters. %DIF = percentage of items manipulated to create DIF. Pattern = difference in manipulated parameters from Group 1. Impact = mean difference from Group 1. P = power. TIE = Type I error.

Table 4
GMH Power and Type I Error Rates – Five-Group Condition

DIF Type	%DIF	Pattern	Impact	Equal *N*s (1000/1000/1000/1000/1000) P	Equal *N*s (1000/1000/1000/1000/1000) TIE	Unequal *N*s (1000/500/500/1000/1000) P	Unequal *N*s (1000/500/500/1000/1000) TIE
–	0	(0/0/0/0/0)	0/0/0/0/0	–	.001	–	.001
			0/−0.5/−0.5/0/0	–	.003	–	.001
a	20	(0/−0.4/0/0/0)	0/0/0/0/0	.178	.002	.132	.003
			0/−0.5/−0.5/0/0	.080	.002	.047	.002
		(0/−0.4/−0.4/0/0)	0/0/0/0/0	.200	.003	.178	.003
			0/−0.5/−0.5/0/0	.172	.003	.100	.001
		(0/−0.4/−0.4/0.4/0.4)	0/0/0/0/0	.243	.004	.240	.003
			0/−0.5/−0.5/0/0	.262	.004	.227	.003
b	20	(0/0.7/0/0/0)	0/0/0/0/0	1.000	.010	.980	.010
			0/−0.5/−0.5/0/0	1.000	.009	.958	.008
		(0/0.7/0.7/0/0)	0/0/0/0/0	1.000	.008	1.000	.009
			0/−0.5/−0.5/0/0	1.000	.009	.998	.011
		(0/0.7/0.7/−0.7/−0.7)	0/0/0/0/0	1.000	.068	1.000	.011
			0/−0.5/−0.5/0/0	1.000	.089	1.000	.018
a/*b*	20	(0/−0.4/0/0/0)	0/0/0/0/0	.882	.009	.847	.011
		(0/0.7/0/0/0)	0/−0.5/−0.5/0/0	.802	.011	.647	.008
		(0/−0.4/−0.4/0/0)	0/0/0/0/0	.928	.013	.898	.010
		(0/0.7/0.7/0/0)	0/−0.5/−0.5/0/0	.835	.008	.815	.010
		(0/−0.4/−0.4/0.4/0.4)	0/0/0/0/0	1.000	.080	1.000	.017
		(0/0.7/0.7/−0.7/−0.7)	0/−0.5/−0.5/0/0	1.000	.048	1.000	.018

Note. DIF Type = manipulation of *a* parameter only, *b* parameter only, or both *a* and *b* parameters. %DIF = percentage of items manipulated to create DIF. Pattern = difference in manipulated parameters from Group 1. Impact = mean difference from Group 1. P = power. TIE = Type I error.

Table 5
MIMIC Power and Type I Error Rates – Three-Group Condition

DIF Type	%DIF	Pattern	Impact	Equal *N*s (1000/1000/1000) P	Equal *N*s (1000/1000/1000) TIE	Unequal *N*s (1000/500/500) P	Unequal *N*s (1000/500/500) TIE
–	0	(0/0/0)	0/0/0	–	.001	–	< .001
			0/−0.5/−0.5	–	.001	–	.001
			0/−0.5/0.5	–	.003	–	.002
a	20	(0/−0.4/0)	0/0/0	.643	.004	.382	.004
			0/−0.5/−0.5	.623	.005	.340	.003
			0/−0.5/0.5	.695	.010	.402	.005
		(0/−0.4/0.4)	0/0/0	.805	.004	.665	.005
			0/−0.5/−0.5	.765	.011	.580	.007
			0/−0.5/0.5	.883	.013	.710	.006
b	20	(0/0.7/0)	0/0/0	.995	.007	.963	.008
			0/−0.5/−0.5	.998	.012	.937	.008
			0/−0.5/0.5	.997	.027	.955	.014
		(0/0.7/−0.7)	0/0/0	1.000	.012	1.000	.011
			0/−0.5/−0.5	1.000	.030	1.000	.012
			0/−0.5/0.5	1.000	.050	1.000	.016
a/*b*	20	(0/−0.4/0)	0/0/0	.998	.005	.980	.006
		(0/0.7/0)	0/−0.5/−0.5	.997	.012	.948	.010
			0/−0.5/0.5	.997	.015	.955	.009
		(0/−0.4/0.4)	0/0/0	1.000	.015	1.000	.005
		(0/0.7/−0.7)	0/−0.5/−0.5	1.000	.016	1.000	.012
			0/−0.5/0.5	1.000	.043	1.000	.017

Note. DIF Type = manipulation of *a* parameter only, *b* parameter only, or both *a* and *b* parameters. %DIF = percentage of items manipulated to create DIF. Pattern = difference in manipulated parameters from Group 1. Impact = mean difference from Group 1. P = power. TIE = Type I error.

Table 6
MIMIC Power and Type I Error Rates – Five-Group Condition

DIF Type	%DIF	Pattern	Impact	Equal *N*s (1000/1000/1000/1000/1000) P	Equal *N*s (1000/1000/1000/1000/1000) TIE	Unequal *N*s (1000/500/500/1000/1000) P	Unequal *N*s (1000/500/500/1000/1000) TIE
–	0	(0/0/0/0/0)	0/0/0/0/0	–	.002	–	.001
			0/−0.5/−0.5/0/0	–	.002	–	.001
a	20	(0/−0.4/0/0/0)	0/0/0/0/0	.610	.008	.395	.003
			0/−0.5/−0.5/0/0	.630	.011	.408	.005
		(0/−0.4/−0.4/0/0)	0/0/0/0/0	.745	.008	.618	.003
			0/−0.5/−0.5/0/0	.798	.012	.622	.006
		(0/−0.4/−0.4/0.4/0.4)	0/0/0/0/0	.952	.008	.868	.008
			0/−0.5/−0.5/0/0	.970	.012	.887	.012
b	20	(0/0.7/0/0/0)	0/0/0/0/0	.998	.010	.955	.010
			0/−0.5/−0.5/0/0	1.000	.020	.947	.017
		(0/0.7/0.7/0/0)	0/0/0/0/0	1.000	.008	.997	.007
			0/−0.5/−0.5/0/0	1.000	.023	.995	.017
		(0/0.7/0.7/−0.7/−0.7)	0/0/0/0/0	1.000	.014	1.000	.017
			0/−0.5/−0.5/0/0	1.000	.040	1.000	.028
a/*b*	20	(0/−0.4/0/0/0)	0/0/0/0/0	1.000	.008	.985	.010
		(0/0.7/0/0/0)	0/−0.5/−0.5/0/0	.997	.019	.950	.011
		(0/−0.4/−0.4/0/0)	0/0/0/0/0	1.000	.009	1.000	.009
		(0/0.7/0.7/0/0)	0/−0.5/−0.5/0/0	1.000	.020	.998	.013
		(0/−0.4/−0.4/0.4/0.4)	0/0/0/0/0	1.000	.011	1.000	.014
		(0/0.7/0.7/−0.7/−0.7)	0/−0.5/−0.5/0/0	1.000	.038	1.000	.020

Note. DIF Type = manipulation of *a* parameter only, *b* parameter only, or both *a* and *b* parameters. %DIF = percentage of items manipulated to create DIF. Pattern = difference in manipulated parameters from Group 1. Impact = mean difference from Group 1. P = power. TIE = Type I error.

Table 7
MG-NCDIF Power and Type I Error Rates – Three-Group Condition

DIF Type	%DIF	Pattern	Impact	Equal *N*s (1000/1000/1000) P	Equal *N*s (1000/1000/1000) TIE	Unequal *N*s (1000/500/500) P	Unequal *N*s (1000/500/500) TIE
–	0	(0/0/0)	0/0/0	–	.06	–	.02
			0/−0.5/−0.5	–	.05	–	.02
			0/−0.5/0.5	–	.07	–	.02
a	20	(0/−0.4/0)	0/0/0	.41	.09	.15	.02
			0/−0.5/−0.5	.38	.10	.17	.04
			0/−0.5/0.5	.41	.11	.17	.03
		(0/−0.4/0.4)	0/0/0	.60	.08	.36	.02
			0/−0.5/−0.5	.57	.09	.39	.04
			0/−0.5/0.5	.62	.11	.40	.03
b	20	(0/0.7/0)	0/0/0	.74	.07	.63	.03
			0/−0.5/−0.5	.69	.08	.64	.04
			0/−0.5/0.5	.76	.08	.64	.04
		(0/0.7/−0.7)	0/0/0	.87	.08	.88	.02
			0/−0.5/−0.5	.85	.08	.84	.03
			0/−0.5/0.5	.87	.09	.85	.06
a/*b*	20	(0/−0.4/0)	0/0/0	.90	.08	.64	.02
		(0/0.7/0)	0/−0.5/−0.5	.82	.09	.66	.03
			0/−0.5/0.5	.78	.08	.66	.03
		(0/−0.4/0.4)	0/0/0	.94	.09	.99	.05
		(0/0.7/−0.7)	0/−0.5/−0.5	1.00	.10	1.00	.03
			0/−0.5/0.5	.99	.10	1.00	.05

Note. The MG-NCDIF results in the power (P) and Type I error (TIE) columns reflect the item parameter replication significance test only. DIF = manipulation of *a* parameter only, *b* parameter only, or both *a* and *b* parameters. %DIF = percentage of items manipulated to create DIF. Pattern = difference in manipulated parameters from Group 1. Impact = mean difference from Group 1. Adapted from "Multiple Group Noncompensatory Differential Item Functioning in Raju's Differential Functioning of Items and Tests," by T. C. Oshima, K. Wright, and N. White, 2015, *International Journal of Testing, 15*, pp. 254-273. Reprinted with permission.

Table 8
MG-NCDIF Power and Type I Error Rates – Five-Group Condition

DIF Type	%DIF	Pattern	Impact	Equal *N*s (1000/1000/1000/1000/1000) P	Equal *N*s (1000/1000/1000/1000/1000) TIE	Unequal *N*s (1000/500/500/1000/1000) P	Unequal *N*s (1000/500/500/1000/1000) TIE
–	0	(0/0/0/0/0)	0/0/0/0/0	–	.07	–	.04
			0/−0.5/−0.5/0/0	–	.07	–	.04
a	20	(0/−0.4/0/0/0)	0/0/0/0/0	.29	.07	.19	.04
			0/−0.5/−0.5/0/0	.27	.07	.18	.05
		(0/−0.4/−0.4/0/0)	0/0/0/0/0	.42	.10	.34	.05
			0/−0.5/−0.5/0/0	.40	.10	.33	.06
		(0/−0.4/−0.4/0.4/0.4)	0/0/0/0/0	.63	.09	.55	.04
			0/−0.5/−0.5/0/0	.63	.12	.53	.04
b	20	(0/0.7/0/0/0)	0/0/0/0/0	.69	.06	.60	.03
			0/−0.5/−0.5/0/0	.68	.08	.63	.04
		(0/0.7/0.7/0/0)	0/0/0/0/0	.79	.06	.71	.04
			0/−0.5/−0.5/0/0	.78	.06	.73	.05
		(0/0.7/0.7/−0.7/−0.7)	0/0/0/0/0	.84	.09	.86	.04
			0/−0.5/−0.5/0/0	.84	.07	.86	.04
a/*b*	20	(0/−0.4/0/0/0)	0/0/0/0/0	.69	.09	.63	.05
		(0/0.7/0/0/0)	0/−0.5/−0.5/0/0	.69	.07	.64	.03
		(0/−0.4/−0.4/0/0)	0/0/0/0/0	.78	.09	.73	.05
		(0/0.7/0.7/0/0)	0/−0.5/−0.5/0/0	.78	.09	.74	.05
		(0/−0.4/−0.4/0.4/0.4)	0/0/0/0/0	1.00	.11	1.00	.06
		(0/0.7/0.7/−0.7/−0.7)	0/−0.5/−0.5/0/0	1.00	.13	1.00	.06

Note. The MG-NCDIF results in the power (P) and Type I error (TIE) columns reflect the item parameter replication significance test only. DIF = manipulation of *a* parameter only, *b* parameter only, or both *a* and *b* parameters. %DIF = percentage of items manipulated to create DIF. Pattern = difference in manipulated parameters from Group 1. Impact = mean difference from Group 1. Adapted from "Multiple Group Noncompensatory Differential Item Functioning in Raju's Differential Functioning of Items and Tests," by T. C. Oshima, K. Wright, and N. White, 2015, *International Journal of Testing, 15*, pp. 254-273. Reprinted with permission.

5 DISCUSSION

This study was designed to compare the performance of the MG-NCDIF index with the GMH and MIMIC DIF detection methods in simulated conditions that modeled both uniform and non-uniform DIF. Additionally, the GMH and MIMIC methods, which have historically used a traditional reference group, were modeled using a base group reference. A critical race theory framework guided this study. The goal was to answer the following questions. First, how does MG-NCDIF perform compared to existing multiple-group DIF detection methods? Second, does the efficacy of the GMH and MIMIC indices vary when detecting various types of DIF (i.e., uniform or non-uniform)?

Overall, the MG-NCDIF method exhibited lower power and higher Type I error rates than the MIMIC method. The MG-NCDIF method did outperform the GMH method when non-uniform DIF was simulated via the *a* parameter only; however, when the *b* parameter was manipulated (to model uniform DIF or non-uniform DIF in combination with manipulation of the *a* parameter), power was higher for the GMH index than the MG-NCDIF index. Across analyses, GMH exhibited lower Type I error rates than MG-NCDIF.

In comparison with the GMH method, the MIMIC method demonstrated higher power when (a) non-uniform DIF via manipulation of the *a* parameter or (b) unidirectional or bidirectional non-uniform DIF via manipulation of the *a* and *b* parameters was present. The two indices performed similarly in regard to power for the remaining conditions. MIMIC exhibited slightly higher Type I error rates than GMH, except in cases when the GMH Type I error rate exceeded .02.

All three methods exhibited higher power for the detection of uniform DIF and non-uniform DIF when both the *a* and *b* parameters were adjusted; power was lower for the detection

of non-uniform DIF when the adjustment was made solely to the *a* parameter. As mentioned earlier, caution must be exercised when comparing the results for different types of DIF because the effect of the type of DIF becomes confounded with the effect of the magnitude of DIF when two parameters are manipulated instead of a single parameter. As expected, the difference in observed performance based on DIF type was most noticeable for the GMH method. Power consistently exceeded .900 for the MIMIC index when uniform DIF or non-uniform DIF via manipulation of the *a* and *b* parameters was modeled and for the GMH index when uniform DIF or bidirectional non-uniform DIF via manipulation of the *a* and *b* parameters was modeled, indicating that the base group is an efficacious reference under these conditions. Given that the base group is likely a representative sample of the full dataset, it is not expected that use of a base group reference would detrimentally impact the remaining conditions, but further study is needed to make such a claim. Therefore, it is suggested that simulations be conducted in which DIF detection rates for traditional and base reference groups are directly compared.

This study is significant for several reasons. First, although a simulation study was conducted to assess the performance of the MG-NCDIF index (Oshima et al., 2015), there had not yet been a simulation study to compare this index with existing methods of DIF detection. Second, most simulation studies that have examined the performance of the GMH and MIMIC methods did so in the context of uniform DIF; there have been few simulation studies that examined the performance of these methods in the context of non-uniform DIF. Finally, the fact that the MG-NCDIF, GMH, and MIMIC indices were modeled using a base (i.e., composite or omnicultural) reference group, the findings of this study contributed to the existing literature base on the use of such a reference group in DIF detection.

The three (multiple-factor) multiple-group DIF detection methods examined in this study all show promising results. Of course, various DIF detection methods may or may not identify the same set of items as exhibiting DIF for the same groups, as Hambleton (2006) expounded:

> it is well-known that competing DIF procedures do not produce identical results and this is not surprising because procedures are themselves quite different – some are model-based, others are not; some condition or match on test score and others on a latent trait or traits, 2-stage DIF remains an option with many of the procedures, and summarizing conditional differences at ability levels is handled differently by the various procedures. (p. S186)

Hambleton's preferred approach to "apply multiple procedures and then especially focus on the items that show the largest statistics with each procedure" (p. S186) may indeed be the most responsible method for practitioners. However, limited resources may preclude such practice, and readers may be interested in the recommendation of a single method. Based on the procedures used in the current study and the results thereof, the following recommendations are provided.

Overall, the MIMIC method exhibited the best performance across the studied conditions, balancing the highest power with acceptable Type I error rates. This index was, by far, the most successful in the identification of non-uniform DIF when only the *a* parameter was manipulated and had exceptional detection rates in the other conditions; its use is therefore recommended when both uniform and non-uniform DIF detection are of interest. If, however, the practitioner is solely interested in the detection of uniform DIF, the GMH method is recommended, based on the excellent power it exhibited and the ease of which this method may be implemented (see below for further discussion). Although the Type I error rates exceeded the nominal alpha level

in some conditions, practitioners may be willing to overlook this downfall, given that the result is that some items that are flagged as exhibiting DIF are, in fact, DIF-free.

Of course, making recommendations based solely on the efficacy of the various approaches (i.e., the power and Type I error results) is a relatively straightforward activity. Ultimately, however, the selection of a DIF detection approach should be based on (a) theoretical and philosophical considerations related to the assessment program and the various candidate DIF detection methods; (b) methodological considerations, such as purification procedures and post-hoc testing; and (c) practical considerations, including requisite sample sizes, time limitations, expense, and the availability of requisite software.

There are several theoretical and philosophical questions related to the assessment program and the various candidate DIF detection methods with which practitioners must wrestle. What is the scoring philosophy of the assessment program, and which DIF method is most consistent with that philosophy? Which matching criterion is preferrable: test scores or a latent trait? What assumptions are associated with each DIF detection method, and are they tenable? Are covariates to be included in the model? The recommendation of a DIF detection method is dependent on the answers to questions such as these. For example, if performance is reported as a raw score for a particular assessment (i.e., IRT procedures such as equating are *not* conducted), the GMH approach would be most consistent with the scoring philosophy. If, however, the assessment program equates forms using a Rasch IRT model, the MIMIC and MG-NCDIF approaches would be more consistent with the operational philosophy. It should be noted that the MG-NCDIF index is the only DIF detection method studied herein that permits modeling of a pseudo-chance (guessing) parameter and is, therefore, the recommended approach when the assessment program employs three-parameter logistic (3PL) IRT modeling. Furthermore, the

GMH index has the advantage of having the fewest assumptions – unidimensionality and a fixed odds ratio across scores – of the studied methods, while MIMIC models assume unidimensionality, local independence, and equal variances for the latent trait across groups and the MG-NCDIF index assumes unidimensionality, local independence, and model fit (Chun et al., 2016; Teresi, 2006). On the other hand, GMH cannot be used to detect DIF when covariates are to be included in the model; in such cases, use of the parametric MIMIC and MG-NCDIF indices would be most appropriate. As a thorough comparison of the assumptions, advantages, and disadvantages associated with these DIF detection methods, as well as other methods, extends beyond the scope of this paper, readers are referred to Hambleton (2006) and Teresi (2006) for an excellent discussion of these topics.

Methodological considerations such as purification procedures and post-hoc testing should also be weighed when evaluating the relative (dis)advantages of various DIF detection approaches. For practitioners and researchers intending to identify a purified anchor set (i.e., conducting a two-stage DIF analysis), the GMH method using the R (R Core Team, 2020) "difR" package (Magis et al., 2010) is both the simplest and quickest of the three methods studied herein. In fact, the "difGMH" function includes an optional argument indicating whether the user wishes to purify the dataset using an iterative procedure, alleviating the need to write additional code. Unfortunately, purification is a more demanding task with the MIMIC and MG-NCDIF indices. For MIMIC, additional M*plus* (Muthén & Muthén, 2007) code is necessitated, and the run time for analyses is nearly doubled. For MG-NCDIF, after identifying a set of anchor items, a second linking process is required and these purified linking coefficients must then be applied to the second stage of the analysis. The MG-NCDIF index does have an advantage over the MIMIC index in that the anchor items are automatically retested for DIF (i.e., full

purification is achieved) without the use of additional coding. Additionally, post-hoc testing may also be of interest to users, to determine which groups differ from the reference group when omnibus significance is found. If post-hoc testing is to be conducted, MG-NCDIF and GMH are the recommended DIF detection approaches of those included in the current study. Recall that MG-NCDIF is defined as

$$\text{MG-NCDIF}_i = \frac{\sum_{g=1}^{p}\sum_{s=1}^{N_B} d_{ig}\left(\hat{\theta}_s\right)^2}{pN_B},$$

where $d_{ig}(\hat{\theta}_s) = P_{iB}(\hat{\theta}_s) - P_{iG_g}(\hat{\theta}_s)$ for group g of p groups. To ascertain which group(s) are exhibiting the greatest difference from the base group, users need only modify the SAS (SAS Institute Inc., 2012) code to print the $\sum_{s=1}^{N_B} d_{ig}\left(\hat{\theta}_s\right)^2$ value for each group while the MG-NCDIF index is calculated; the higher this value is for a group, the more that group is contributing to the significant omnibus DIF result. In other words, to alleviate the need for post-hoc tests, the code may be written to show pairwise NCDIF values prior to combining them in a single MG-NCDIF index. Post-hoc testing is also quite simple for the GMH index: simply conduct pairwise analyses of each group against the base group using the MH method (Finch, 2016) or MH with a Bonferroni-adjusted alpha level (Penfield, 2001). Conversely, post-hoc testing with the MIMIC index is both complex and time-consuming; it necessitates additional experimental model comparisons in which model paths are freed or dropped, an endeavor that has yet to be studied by DIF scholars and which would entail substantial additional coding (either in R or M*plus*).

Practical considerations invariably come into play when selecting a DIF detection method. In authentic testing situations, psychometricians are often expected to turn around student scores in an expedited timeframe. Of the methods featured in the current study, the GMH

approach is likely the most practical option for many assessment companies and researchers. As a non-parametric method, it has the lowest sample size requirements of the three approaches, a desirable trait when planning field testing and when meeting tight reporting deadlines for operational field test assessment items. It is also the most time-efficient DIF detection method, as it requires writing only a small amount of R code and the two-stage analysis of a 30-item assessment takes approximately one minute. As added benefits, the M*plus* code is short and simple, and the hand-calculated χ^2 difference tests of nested models are easy to conduct. The availability of R at no cost to the user is also a relative advantage for this method. The parametric approaches, on the other hand, have higher sample size requirements, take considerably more time to complete, require more coding expertise, and have higher associated software licensing fees. As explained earlier, for the MIMIC method, nested models must be compared separately for each studied item; for a 30-item assessment, this could take several hours. The process may be automated using R, as was done herein, which would reduce the run-time significantly, but it requires the practitioner to know an additional coding language. The fees for M*plus* licensing may also be prohibitive for researchers or small organizations. Similarly, due to the necessity of linking each focal group to the reference group, the MG-NCDIF approach is more time consuming and requires multiple software applications, and these licenses may be costly as well. Furthermore, the "DIFCUT" program (Nanda, Oshima, & Gagne, 2005), which is currently available only in SAS (SAS Institute Inc., 2012), must be modified based on the number of groups in a particular study, an endeavor that requires extra time and coding ability.

One final recommendation merits attention: the selection of a reference group. Each of the three DIF detection methods studied herein may be conducted with a base reference group, despite the fact that GMH and MIMIC studies have historically been conducted with a traditional

reference group. It is worth noting that, although the use of a base group is recommended when identifying DIF between racial/ethnic groups, its use may not always be appropriate. For example, in the case of a test designed to assess students' command of the English language, the use of a traditional reference group comprised of native English speakers is a defensible decision because, ultimately, it is desirable for the performance of English learners to be commensurate with the performance of students for whom English is their first language and for items to function identically between these two groups. Therefore, it is strongly recommended that, as discussed earlier, the choice of a reference group be based on the definitions of fairness and bias that are most applicable to the research study being conducted or the assessment program goals.

At this point, it is typical for the researcher to identify a few major limitations of his/her study and recommend future quantitative endeavors. It is here that I wish to rejoin my readers in a more personal manner. As I discussed earlier, CRT and QuantCrit scholars advocate for de-anonymized research in which the researcher is forthcoming about decisions related to analytical design, data, interpretations, and analysis. I have attempted to provide as much transparency as possible regarding the methodological decisions that I made throughout the course of this study because, as Garcia and Mayorga (2018) eloquently explained,

> exogenous forces require that researchers make compromises in their research practices, yet it is important to note that these decisions are not viewed as compromises, but normal decisions made in the course of conducting research . . . these small, 'normal' compromising decisions are where white supremacy is reproduced and comes to bear through research outcomes. (p. 246)

Herein, I have acknowledged the compromises that I made in the course of my work, making myself professionally vulnerable in the process. Each compromise that I made during this

research study became an inherent limitation to the study. For example, in choosing to manipulate the *a* parameter by 0.4, my study is limited by the fact that I didn't explore smaller differences in each group's *a* parameter, such as 0.2, or larger differences, such as 0.6. In the interest of providing guidance to my readers, and to adhere to the tradition of naming just a few limitations, I am happy to identify those that I believe to be the most important. (Of course, you are welcome to respectfully disagree.)

With one exception, which I will discuss momentarily, the most significant limitations to my study shared a common cause, that most precious and limited resource: time. First, the analyses conducted in the current study did not include DIF detection with a traditional reference group, which would have allowed for a direct comparison of the performance of the base reference group against the traditional reference group for each DIF detection method. The results herein indicate that use of a base reference group is potentially as efficacious as the use of a traditional reference group. If that turns out to be true, it provides quantitative justification for a change in the psychometric status quo, in addition to the other justifications provided earlier in this study. In other words, such studies could provide a more solid statistical foundation on which to build a movement to end the "Whiteness as the ideal to be reached" mentality in testing. It is, therefore, imperative that these analyses be conducted in the future. The second limitation is that the performance of the MFMG-NCDIF and MIMIC methods were not compared when multiple background factors (i.e., covariates), each with varying numbers of groups, were simulated. Third, unlike the procedures for the MG-NCDIF and GMH indices, in which each anchor item was retested for DIF against the remaining anchor items, only partial purification (Fikis & Oshima, 2017) was carried out for the MIMIC index. Fourth, no post-hoc testing was conducted, a shortcoming given that GMH and MIMIC are omnibus methods that signal the

identification of DIF for at least one pairwise group comparison. Fifth, the performance of the MG-NCDIF method was not re-analyzed using the datasets from the current study.

The final limitation that I would like to acknowledge, which was related not to time but to DIF detection methods in general, is that of anchor selection. As readers likely noted, the number of anchors may be arbitrarily selected in a variety of ways. On one end of the spectrum, quantitative researchers may limit the number of anchors to a set maximum (e.g., five items), as I did with the MIMIC method, to limit the chances of having a contaminated anchor set. At the other end of the spectrum is the option to allow as many anchors as possible to make use of as much "clean" data as possible, which researchers may prefer in studies in which the MH and/or GMH methods are used, given that the total (matching) score, comprised solely of the anchor items and studied item, is a key part of these analyses. Ultimately, I decided to limit the number of anchor items in my MIMIC analyses because there was a clear precedent for this in the literature. On the other hand, there was no evidence of this practice in the GMH literature; thus, although the risk of anchor contamination was greater, I allowed anchor items to be identified without any upper limit. Neither anchor-count method is inherently "correct"; the compromise here was that it resulted in an apples-to-oranges comparison of DIF indices. Furthermore, the subjective decisions around anchoring became even more complex when I had to choose whether to use a pre-defined set of anchors or a set of anchors identified in the first stage of DIF detection procedures. It was my perspective that it was best if I refrained from selecting the anchors myself (except when absolutely necessary, as discussed earlier) because, in an authentic assessment situation, it is unlikely that the researcher will know which items are truly DIF-free; (s)he will have to rely on the statistical procedure(s) being implemented. When it was necessary for me to intervene in the identification of MIMIC anchors, I attempted to do so conservatively, making

use of a single highly-discriminating item, a practice supported in the literature (e.g., Chun et al., 2016). I acknowledge that other quantitative scholars may have made different choices.

It should come as no surprise, therefore, that my list of recommendations for future research would include study of (1) the performance of the base reference group in direct comparison with a traditional reference group, (2) the performance of the MFMG-NCDIF and MIMIC methods with multiple background factors (i.e., covariates), (3) MIMIC performance with full anchor purification, and (4) GMH and MIMIC post-hoc testing, where appropriate.

Rios-Aguilar (2014) asserted that "a more open discussion of the decisions made, drawbacks, and surprises while scholars do research is certainly needed if we aspire to conduct research that matters in producing equitable opportunities for all students" (p. 97), and it was my sincere intent to do so. However, from a critical perspective it is not sufficient to be transparent in our decision making as researchers; it is also our responsibility to think critically about the analytical design, data, and interpretations of our studies. Although this was a simulation study, I consider it to be a racial project; consequently, I feel it would be negligent not to trouble DIF analyses and provide recommendations for readers. Therefore, I humbly offer three recommendations to those who study DIF methods or make use of them in authentic assessment contexts.

First, critically examine the classifications that are embedded in DIF studies, particularly those corresponding to race and ethnicity. In echo of Gillborn et al. (2018), "categories are neither 'natural' nor given" (p. 169). Race is a social construct, not an objective, scientific, biological characteristic as it is often treated (Ladson-Billings, 2009). Furthermore, racial categorization can obscure intra-group heterogeneity (Garcia & Mayorga, 2018; Teranishi, 2007). As Pérez Huber et al. (2018) explained, "aggregate data . . . can mask patterns of racial

inequity that become prominent when data is disaggregated by ethnic subgroup, language, geography, immigration status, and other demographic variables" (p. 227). Garcia and Mayorga (2018) offered a related piece of advice: "acknowledge the limitations of generalizability of sample when speaking of racial groups as homogeneous populations/categorizations, especially when you recode samples to create monologic groups for statistical purposes (e.g. Non-white . . .)" (p. 249). We, as a quantitative field, need to move beyond the typical "White, Black, Asian, Hispanic, American Indian, or Multiracial" categorization to develop a deeper understanding of how items function between – and within – groups of examinees.

My second recommendation is to be exceedingly cautious when discussing examinee ability, or theta (θ). As many critical scholars have outlined in detail, quantitative analyses, statistics, and assessment – outgrowths of intelligence testing – have racist origins that may be traced to early eugenics movements (e.g., Gould, 1996; López et al., 2018). Historically, quantitative studies have controlled for variables affected by institutional racism, such as "intelligence" and prior achievement, and such studies have fed the deficit discourse surrounding students of color. [*The Bell Curve* (Herrnstein & Murray, 1994) is a particularly egregious example of this type of work.] As Sleeter (2004) explained, during the late 1980s, "most school reforms that were discussed emphasized raising standards and requiring students to work harder, and the 'at risk' discourse emerged to describe those who were falling behind (who were mainly children of color and children from low-income backgrounds)" (p. 166). This discourse continues today. As Ladson-Billings (2009) discussed, "current instructional strategies presume that African American students are deficient. As a consequence, classroom teachers are engaged in a never-ending quest for 'the right strategy or technique' to deal with (read: control) 'at risk' (read: African American) students" (pp. 29-30). In the case of works such as *The Bell Curve*

(Herrnstein & Murray, 1994), studies that document alleged student deficiencies have been aggregated to make a case for group deficiencies. In other words, there is a tragic cycle in which students are victimized by a racist education system, this racism is statistically factored out of achievement and ability calculations, these same students are then portrayed as having lesser potential, and this data is then translated into discourses of racial patterns of ineptitude.

López et al. (2018) asserted that we must "mov[e] away from erroneous genetic or cultural essentialist logics that conceptualize differences in intelligence and academic performance as innate and unchanging" (p. 200). Today, items are identified as exhibiting DIF if they function differently for examinees of equal ability from different groups. "Ability" in the DIF context, I fear, may be misunderstood as "intelligence." As presented earlier, in the IRT framework, the probability of an examinee correctly answering an assessment item *i*, denoted as $P_i(\theta)$, is given by the three-parameter logistic (3PL) function

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}},$$

where a_i represents the discrimination parameter for item *i*, b_i is the difficulty parameter for item *i*, c_i is the pseudo-guessing parameter for item *i*, θ represents examinee ability, and *D* is the scaling constant of 1.7. I argue that, conceptually, this function should be rewritten as

$$P_i(\theta_t) = c_i + (1 - c_i)\frac{e^{Da_i(\theta_t - b_i)}}{1 + e^{Da_i(\theta_t - b_i)}}$$

where θ_t represents examinee ability at time *t*. In layman's terms, performance is dependent on an examinee's ability *at a particular time*. Of course, I have no expectation that the formal mathematical model will be changed. My intent here is to emphasize that, in referencing ability, scholars have the responsibility to communicate the message that it is not innate or static; ability is a dynamic trait that, like academic achievement, reflects "different[ial] treatment,

opportunities, and exposure to structural, institutional, and interpersonal racism" (López et al., 2018, p. 193).

My third recommendation, closely related to the previous one, is to critically revisit the (ab)uses to which we put DIF results, especially those associated with authentic assessments. Currently, DIF results are used to determine which items should be revised or omitted entirely from assessments, but not to trouble the educational system itself. When an item is found to function differentially, psychometricians and content specialists may ask "What characteristics of this item caused it to exhibit DIF?" However, we should also ask, "What characteristics of instruction and/or the educational system caused this item to exhibit DIF?" Furthermore, discussions of group mean differences in ability are particularly susceptible to misinterpretation and must be explicitly defined and contextualized. As with all other quantitative studies, findings must be discussed within – not isolated from – the sociohistorical context (i.e., the lived experiences of students of color), because "opportunity structures, not innate, genetic, or cultural differences, shape the contours of . . . education outcomes and accompanying sedimented inequalities" (López et al., 2018, p. 188). Consider, for example, the following scenario . . . .

Michelle Darden walked around the room, monitoring her 4th grade students as they worked in near silence on their statewide end-of-year standardized assessment. She wished she could tell how they were doing; she wanted them to feel successful; she wanted them to recognize how much they had learned – not just from her, but from each other – and she wanted their parents to see it too.

As Michelle began another loop around the room, Dymond, one of many Black students in the class, raised her hand. "Ms. Darden, number 25 says there are two correct answers, but I see three." Michelle read the test item to which Dymond was pointing.

Source 2: Some Ways to Be a Good Citizen

Source 3: Thomas Jefferson on Citizenship (1792)

Thomas Jefferson was the main author of the Declaration of Independence and the third president of the United States. At the time of this quotation, he was serving as secretary of state under President George Washington.

"A nation, as a society, forms a moral person, and every member of it is personally responsible for his society."

Based on Source 2 and Source 3, which statements **best** describe the qualities of good citizens? Select the **two** correct answers.

A. They go to college.
B. They use their talents.
C. They spend money.
D. They vote in elections.
E. They volunteer their time.
F. They study history.

It was obvious to Michelle that the author's intent was for students to identify options D and E. However, she was fairly sure she understood why the young girl thought there were three correct answers. Dymond's family had moved to Louisiana a few years ago from Alabama,

where her great-grandfather had participated in the Civil Rights movement. They had been fortunate that he was able to speak to the class, and the students had been captivated as he recounted his experiences. Michelle had been equally enthralled. His story brought meaning to a movement she had never fully understood as a White woman; it grounded the message that nearly every one of her social studies teachers had told her: You need to study history so that you don't repeat it. To Dymond and her Black peers, studying history was an essential step to understanding racism and the foundation for continuing her great-grandfather's pursuit of justice and equity for people of color.

Michelle felt helpless, thinking that Dymond surely viewed the study of history as something in which a good citizen would engage, yet recognizing that her professional obligations did not permit her to offer any help. She gave Dymond a soft squeeze on the shoulder, replying in a whisper, "I'm not allowed to help you with that. Just do the best you can. You've got this!"

Five minutes later, walking past Rodrigo, she took note of the puzzled look etched upon his face. "Everything okay?" she asked.

"I don't know what to do. The computer will only let me choose two answers." He nodded towards the screen, where number 25 was displayed.

"Oh, no!" Michelle thought, "Not again!"

Rodrigo's parents had attended every single parent-teacher conference this year, despite the fact that they worked long hours and that each meeting took twice as much time as other conferences, owing to the need for a Spanish translator. Even though Rodrigo was only 10 years old, they wanted to make sure that he was mastering the state standards so that he could attend the magnet school across town when he started 9th grade. It was their hope that he earn a

college degree. His papelito *would symbolize the sacrifices that his extended family had made and the fact that they had established roots in the U.S.; it would also enable him to establish a career in which he would serve the Latina/o community.*

But she could do nothing more in the moment than reply, "You can only choose two answers. Do your best and don't worry, it's only one question."

Michelle began to get worried; if Dymond and Rodrigo each found value in an answer she knew to be considered incorrect, they were surely not the only ones. And, as though she had commanded it, a hand rose into the air at the back of the room. With a quiet turn, Michelle headed towards Graciela. Arriving at her pupil's desk, she again saw number 25 on the monitor.

"Ms. Darden, I know D is right, but I can't decide between C and E for the second answer. When people got COVID and my tía lost her job, she said it was because people didn't have enough money to spend eating out. Doesn't it help the community when you shop and eat at restaurants? Then people have jobs."

Michelle was stunned. She debated how to answer Graciela in a way that would honor her thought process, even as she knew that she was not permitted to offer help and that, if Graciela chose C, she would be wrong in the eyes of the test developer and the state. In the most supportive tone she could muster, she told Graciela how proud she was of the girl's careful thinking about the question and reminded her that she could always mark two answers and come back to the question later if she wanted to think about it more.

Walking back to the front of the room, Michelle sent up a silent prayer. "Please, let this be a field test item."

Although this story is fictional, the test item is not. It was taken from Louisiana's "LEAP 2025 Annotated Social Studies Practice Test Items" guide (Louisiana Department of Education,

2020, pp. 15 and 17). To the author of the test item, a good citizen is defined as a person who is an active participant in democratic processes and who gives his time to the community without expectation of payment. To a Black student like Dymond, a good citizen is defined as a person who learns about the processes that led to and continue to support the marginalization of particular groups of people in our society. To a Latino student like Rodrigo, a good citizen is someone who values education and takes advantage of the educational opportunities afforded him. To a student from a low-income household like Graciela, a good citizen may contribute to the economy by purchasing goods and retaining the services of others.

From a critical race theory framework, we – test developers, psychometricians, educators, and researchers – are ethically bound to ask the questions "What knowledge counts? Whose knowledge counts?" When an item is identified as exhibiting DIF, revising the item or discarding it must not be the only two courses of action. Instead, we must critically examine that item and interrogate the role it plays in the marginalization of the Other, acknowledging that student assessment is simultaneously a reflection of the inequitable instruction that students have received and a racial project that reifies deficit perceptions of children of color as less intelligent, unmotivated, or lacking familial support. It is only by critically examining differentially-functioning items that we will begin to turn this particular *racial project* into a *racial justice project*.

REFERENCES

Asparouhov, T., & Muthén, B. O. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling*, *21*(4), 1-14.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, *57*(1), 289-300.

Buddin, R., McCaffrey, D. F., Kirby, S. N., & Xia, N. (2007). *Merit pay for Florida teachers* (Working Paper No: WR-508-FEA). Santa Monica, CA: RAND Corporation.

Buras, K. L. (2014). From Carter G. Woodson to critical race curriculum studies: Fieldnotes on confronting the history of white supremacy in educational knowledge and practice. In A. D. Dixson (Ed.), *Researching race in education: Policy, practice and qualitative research*. Charlotte, NC: Information Age Publishing.

Chun, S., Stark, S., Kim, E. S., & Chernyshenko, O. S. (2016). MIMIC methods for detecting DIF among multiple groups: Exploring a new sequential-free baseline procedure. *Applied Psychological Measurement*, *40*(7), 486-499.

Clauser, B. E., & Mazor, K. M. (1998). An NCME instructional module on using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*(1), 31-44.

Colby, S. L., & Ortman, J. M. (2015). Projections of the size and composition of the US population: 2014 to 2060. Retrieved from: https://www.census.gov/content/dam/Census/library/publications/2015/demo/p25-1143

Covarrubias, A., Nava, P. E., Lara, A., Burciaga, R., Vélez, V. N., & Solórzano, D. G. (2018). Critical race quantitative intersections: A *testimonio* analysis. *Race, Ethnicity, and Education*, *21*(2), 253-273.

Crenshaw, K., Gotanda, N., Peller, G., & Thomas, K. (1995). Introduction. In K. Crenshaw, N. Gotanda, G. Peller, & K. Thomas (Eds.), *Critical race theory: The key writings that formed the movement* (pp. xiii-xxxii). New York: The New Press.

DeCuir, J. T., & Dixson, A. D. (2004). "So when it comes out, they aren't that surprised that it is there": Using critical race theory as a tool of analysis of race and racism in education. *Educational Researcher*, *33*(5), 26-31.

Dell-Ross, T. L., Oshima, T. C., & Wright, K. D. (2017, April). Demonstration of multiple-factor multiple-group non-compensatory DIF. Paper presented at the annual meeting of the National Council on Measurement in Education, San Antonio, TX.

Ellis, B. B., & Kimmel, H. D. (1992). Identification of unique cultural response patterns by means of item response theory. *Journal of Applied Psychology*, *77*(2), 177-184.

Fidalgo, Á. M. (2011). GMHDIF: A computer program for detecting DIF in dichotomous and polytomous items using generalized Mantel-Haenszel statistics. *Applied Psychological Measurement*, *35*(3), 247-249.

Fidalgo, Á. M., & Madeira, J. M. (2008). Generalized Mantel-Haenszel methods for differential item functioning detection. *Educational and Psychological Measurement*, *68*(6), 940-958.

Fidalgo, Á. M., & Scalon, J. D. (2010). Using generalized Mantel-Haenszel statistics to assess DIF among multiple groups. *Journal of Psychoeducational Assessment*, *28*(1), 60-69.

Fikis, D. R. J., & Oshima, T. C. (2017). Effect of purification procedures on DIF analysis in IRTPRO. *Educational and Psychological Measurement*, *77*(3), 415-428.

Finch, W. 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.

Finch, W. H. (2016). Detection of differential item functioning for more than two groups: A Monte Carlo comparison of methods. *Applied Measurement in Education*, *29*(1), 30-45.

Finch, W. H., & French, B. F. (2019). *Educational and psychological measurement*. New York: Taylor & Francis.

Fleishman, J. A., Spector, W. D., & Altman, B. M. (2002). Impact of differential item functioning on age and gender differences in functional disability. *Journal of Gerontology: Social Sciences*, *57B*(5), S275-S283.

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

Garcia, N. M., & Mayorga, O. J. (2018). The threat of unexamined secondary data: A critical race transformative convergent mixed methods. *Race, Ethnicity, and Education*, *21*(2), 231-252.

Gillborn, D. (2009). Education policy as an act of white supremacy: Whiteness, critical race theory and education reform. In E. Taylor, D. Gillborn, & G. Ladson-Billings (Eds.), *Foundations of critical race theory in education* (pp. 51-69). New York: Routledge.

Gillborn, D. (2010). The colour of numbers: Surveys, statistics and deficit-thinking about race and class. *Journal of Education Policy*, *25*(2), 253-276.

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.

Gould, S. J. (1996). *The mismeasure of man*. New York: W. W. Norton & Company, Inc.

Hallquist, M. N., & Wiley, J. F. (2018). *MplusAutomation*: An R package for facilitating large-scale latent variable analyses in M*plus*. *Structural Equation Modeling, 25*, 621-638. doi: 10.1080/10705511.2017.1402334.

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical Care, 44*(11), S182-S188.

Hassel, B. C., & Hassel, E. A. (2007). *Improving teaching through pay for contribution*. Washington, DC: National Governors Association Center for Best Practices.

Herrnstein, R. J., & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life*. New York: Simon and Schuster, Inc.

Huggins, A. C., & Penfield, R. D. (2012). An NCME instructional module on population invariance in linking and equating. *Educational Measurement: Issues and Practice, 31*(1), 27-40.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association, 70*(351), 631-639.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling, 15,* 136-153.

Kim, J., & Oshima, T. C. (2012). Effect of multiple testing adjustment in differential item functioning detection. *Educational and Psychological Measurement, 75*(3), 458-470.

Klein, A., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika, 65*(4), 457-474.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York: Guilford Press.

Ladson-Billings, G. (2004). Landing on the wrong note: The price we paid for *Brown*. *Educational Researcher, 33*(7), 3-13.

Ladson-Billings, G. (2009). Just what is critical race theory and what's it doing in a *nice* field like education? In E. Taylor, D. Gillborn, & G. Ladson-Billings (Eds.), *Foundations of critical race theory in education* (pp. 51-69). New York: Routledge.

Ladson-Billings, G., & Tate, W. (2006). Toward a critical race theory in education. In A. D. Dixson & C. K. Rousseau (Eds.), *Critical race theory in education: All God's children got a song* (pp. 11-30). New York: Routledge.

Landis, J. R., Heyman, E. R., & Koch, G. G. (1978). Average partial association in three-way contingency tables: A review and discussion of alternative tests. *International Statistical Review, 46*, 237-254.

López, N., Erwin, C., Binder, M., & Chavez, M. J. (2018). Making the invisible visible: Advancing quantitative methods in higher education using critical race theory and intersectionality. *Race, Ethnicity, and Education, 21*(2), 180-207.

Lopez Rivas, G. E., Stark, S., & Chernyshenko, O. S. (2009). The effects of referent item parameters on differential item functioning detection using the free baseline likelihood ratio test. *Applied Psychological Measurement, 33*(4), 251-265.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Louisiana Department of Education. (2020, December 21). LEAP 2025 Annotated Social Studies Practice Test Items. Retrieved from https://www.louisianabelieves.com/docs/default-source/assessment/leap-2025-annotated-social-studies-practice-test-items.pdf?sfvrsn=8

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*(3), 847-862.

Magis, D., Raîche, G., Béland, S., & Gérard, P. (2011). A generalized logistic regression procedure to detect differential item functioning among multiple groups. *Internal Journal of Testing*, *11*, 365-386.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719-748.

Mayhew, M. J., & Simonoff, J. S. (2015). Non-White, no more: Effect coding as an alternative to dummy coding with implications for higher education researchers. *Journal of College Student Development*, *56*(2), 170-175.

Meade, A. W., & Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. *Journal of Applied Psychology*, *97*(5), 1016–1031.

Morris, J. (2006). The forgotten voices of black educators. In A. D. Dixson & C. K. Rousseau (Eds.), *Critical race theory in education: All God's children got a song* (pp. 129-151). New York: Routledge.

Muthén, L. K. (2009, July 1). Standardized coefficients and fit indices with MLR [Online forum comment]. Retrieved from http://www.statmodel.com/cgi-bin/discus/discus.cgi?pg=prev&topic=11&page=4464

Muthén, L. K., & Muthén, B. O. (2007). M*plus*: Statistical Analysis with Latent Variables (Version 8.2) [Computer software]. Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2017). M*plus* User's Guide (8th ed.). Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (n.d.). Chi-square difference testing using the Satorra-Bentler scaled chi-square. Retrieved from https://www.statmodel.com/chidiff.shtml

Muthén, B. O. (2004, July). *General latent variable modeling using Mplus Version 3: Block 1: Structural equation modeling*. Workshop presented at the meeting of Society for Multivariate Analysis in the Behavioral Sciences (SMABS), Jena, Germany. Retrieved from https://www.statmodel.com/download/MuthenMplusWorkshop1.pdf

Nanda, A. O., Oshima, T. C., & Gagne, P. (2005). DIFCUT: A SAS-IML program for conducting significance tests for differential functioning of items and tests (DFIT). *Applied Psychological Measurement, 30*, 150-151.

Omi, M., & Winant, H. (1994). *Racial formation in the United States*. New York: Routledge.

Oshima, T. C., & Morris, S. B. (2008). An NCME instructional module on Raju's Differential Functioning of Items and Tests (DFIT). *Educational Measurement: Issues and Practice, 27*, 43-50.

Oshima, T. C., Raju, N. S., & Nanda, A. O. (2006). A new method for assessing statistical significance in the differential functioning of items and tests (DFIT) framework. *Journal of Educational Measurement, 43*(1), 1-17.

Oshima, T. C., Wright, K., & White, N. (2015). Multiple-group noncompensatory differential item functioning in Raju's Differential Functioning of Items and Tests. *International Journal of Testing, 15*, 254-273.

Pendergast, L. L., von der Embse, N., Kilgus, S. P., & Eklund, K. R. (2017). Measurement equivalence: A non-technical primer on categorical multi-group confirmatory factor analysis in school psychology. *Journal of School Psychology, 60*, 65-82.

Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: A comparison of three Mantel-Haenszel procedures. *Applied Measurement in Education, 14*(3), 235-259.

Pérez Huber, L., Vélez, V. N., & Solórzano, D. (2018). More than 'papelitos': A QuantCrit counterstory to critique Latina/o degree value and occupational prestige. *Race, Ethnicity, and Education*, *21*(2), 208-230.

R Core Team (2020). R: A language and environment for statistical computing (Version 4.0.3) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, *19*(4), 353-368.

Rebouças, D. A., & Cheng, Y. (2019). Relationship between item characteristics and detection of differential item functioning under the MIMIC model. *Psychological Test and Assessment Modeling*, *61*(2), 227-257.

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

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*(2), 105-116.

RStudio Team (2020). RStudio: Integrated development environment for R (Version 1.3.1093) [Computer software]. Boston, MA: RStudio, PBC. Retrieved from http://www.rstudio.com/

Sari, H. I., & Huggins, A. C. (2015). Differential item functioning detection across two methods of defining group comparisons: Pairwise and composite group comparisons. *Educational and Psychological Measurement*, *75*(4), 648-676.

SAS Institute, Inc. (2012). SAS [Computer software]. Cary, NC: Author.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*(4), 507-514.

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika*, *75*(2), 243-248.

Seybert, J., & Stark, S. (2012). Iterative linking with the differential functioning of items and tests (DFIT) method: Comparison of testwide and item parameter replication (IPR) critical values. *Applied Psychological Measurement*, *36*, 494-515.

Shealy, R. T., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58,* 159-194.

Shih, C.-L., & Wang, W.-C. (2009). Differential item functioning detection using the multiple indicators, multiple causes method with a pure short anchor. *Applied Psychological Measurement*, *33*(3), 184-199.

Sireci, S. G., & Rios, J. A. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation*, *19*(2-3), 170-187.

Sleeter, C. E. (2004). How white teachers construct race. In G. Ladson-Billings & D. Gillborn (Eds.), *The RoutledgeFalmer reader in multicultural education* (pp. 163-178). New York: RoutledgeFalmer.

Solórzano, D., & Yosso, T. (2009). Critical race methodology: Counter-storytelling as an analytical framework for education research. In E. Taylor, D. Gillborn, & G. Ladson-Billings (Eds.), *Foundations of critical race theory in education* (pp. 131-147). New York: Routledge.

Somes, G. W. (1986). The generalized Mantel-Haenszel statistic. *The American Statistician, 40*(2), 106-108.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201-210.

Sullivan, E., Larke, P. J., & Webb-Hasan, G. (2010). Using critical policy and critical race theory to examine Texas' school disciplinary policies. *Race, Gender & Class*, *17*(1/2), 72-87.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361-370.

Taylor, E. (2009). The foundations of Critical Race Theory in education: An introduction. In E. Taylor, D. Gillborn, & G. Ladson-Billings (Eds.), *Foundations of critical race theory in education* (pp. 1-13). New York: Routledge.

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

Teresi, J. A. (2006). Different approaches to differential item functioning in health applications: Advantages, disadvantages and some neglected topics. *Medical Care*, *44*(11), S152-S170.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99*, 118-128.

UCLA Statistical Consulting Group. (n.d.). How can I compute a chi-square test for nested models with the MLR or MLM estimators? Mplus FAQ. Retrieved from https://stats.idre.ucla.edu/mplus/faq/how-can-i-compute-a-chi-square-test-for-nested-models-with-the-mlr-or-mlm-estimators/

Wang, W.-C., & Shih, C.-L. (2010). MIMIC methods for assessing differential item functioning in polytomous items. *Applied Psychological Measurement*, *34*(3), 166-180.

Wang, W.-C., & Su, Y.-H. (2004). Factors influencing the Mantel and generalized Mantel-Haenszel methods for the assessment of differential item functioning in polytomous items. *Applied Psychological Measurement, 28*(6), 450-480.

Woods, C. M. (2009a). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, *33*(1), 42–57.

Woods, C. M. (2009b). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research*, *44*, 1-27.

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement*, *35*, 339-361.

Woods, C. M., Oltmanns, T. F., & Turkheimer, E. (2009). Illustration of MIMIC-model DIF testing with the Schedule for Nonadaptive and Adaptive Personality. *Journal of Psychopathology and Behavioral Assessment*, *31*(4), 320-330.

Wright, K. D., & Oshima, T. C. (2015). An effect size measure for Raju's differential functioning for items and tests. *Educational and Psychological Measurement, 75*(2), 338-358.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). BILOG-MG3 (Version 3.0) [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, *30*(3), 233-251.