PATTON, ELIZABETH ADELE, Ph.D. Comparing Three Multilevel Frameworks for the Detection of Differential Item Functioning. (2019)
Directed by Drs. Randall Penfield and Robert Henson. 254 pp.

Multilevel data complicates the accumulation of validation evidence. Using a unilevel approach to differential item functioning in the presence of multilevel data is both a theoretically and statistically unsound method. This simulation study compares three multilevel frameworks for the detection of differential item functioning. The methods compared were the Beggs Mantel-Haenszel adjustment, the multilevel Rasch model, and the SIBTEST bootstrapped standard error adjustment. Five conditions were varied in this study: the magnitude of DIF, the social-unit level sample size, the presence of impact, the degree of correlation within clusters, and the ratio of the reference to focal group. The results suggest that the Beggs Mantel-Haenszel adjustment is superior when analyzing Type I error and power rates. However, the multilevel Rasch model produced more accurate and precise estimates of effect size. Additionally, the multilevel Rasch model has the potential to provide more nuanced information regarding the causes of item bias.

COMPARING THREE MULTILEVEL FRAMEWORKS FOR THE DETECTION OF DIFFERENTIAL ITEM FUNCTIONING

by

Elizabeth Adele Patton

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

Greensboro
2019

Approved by

____
Committee Co-Chair

____
Committee Co-Chair

APPROVAL PAGE

This dissertation, written by Elizabeth Adele Patton, has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Co-Chair

Committee Co-Chair

Committee Members

Date of Acceptance by Committee

Date of Final Oral Examination

ACKNOWLEDGMENTS

I would like to thank my dissertation committee of Dr. Micheline Chalhoub-Deville, Dr. Ric Luecht, Dr. Bob Henson, and Dr. Randy Penfield for their support and feedback throughout this process. In addition, I would like to thank Dr. Micheline Chalhoub-Deville for the significant mentorship she has offered over the years. Her personal and career advice has proven invaluable. Lastly, I would be remiss if I did not thank Dr. Micheline Chalhoub-Deville, Dr. Ric Luecht, and Dr. John Willse for the numerous professional recommendation letters they have provided. I am fortunate to have the endorsement of such well respected professionals and grateful for their assistance in securing a wonderful job. Lastly, I would like to thank Dr. Feifei Ye at the University of Pittsburgh for recognizing a Sociology of Education student who was slightly out of place in her advanced quantitative methods courses. Her guidance led me down this path and to the University of North Carolina Greensboro.

I would like to thank Lexi Lay for her friendship during this process. She is an incredible friend and colleague and I am excited to see what these next stages of life bring for both of us personally and professionally. Finally, and most importantly, I would like to thank my family for their continued love and support through this process. My husband made significant sacrifices, including relocating and career transitions, for me to pursue this degree. My daughter, Felicity, has provided the impetus for much-needed study breaks and is a constant reminder of how important a work-life balance is.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Fairness in testing is a pivotal though complex issue. According to the *Standards for Educational and Psychological Testing* (2014; referred to hereafter as the *Standards*), fairness, “has no single technical meaning and is used in many different ways in public discourse” (p. 49). However, definitions related to testing frequently associate issues of fairness and validity. For instance, ETS (2014) defines fairness as, “the extent to which the inferences made on the basis of test scores are valid for different groups of test takers” (p. 19). While it is not feasible to investigate fairness for all groups in the population of test takers, testing programs should investigate fairness for those groups that experience or research has indicated are likely to be adversely impacted by construct-irrelevant influences on their test performance (ETS, 2014). Typically, analyses include groups which have been discriminated against based on ethnicity, disability status, gender, native language, or race and are defined legally.

The glossary of the *Standards* offers a related definition stating that, “fairness minimizes the construct-irrelevant variance associated with individual characteristics and testing contexts that otherwise would compromise the validity of scores for some individuals” (2014, p. 219). The chapter dedicated to fairness within the *Standards* goes on to describe fairness in testing as being attributable to four principles: (1) fair and equitable treatment of all test takers during the testing process, (2) the lack or absence of measurement bias, (3) access to the constructs measured, and (4) fairness as validity of individual test score interpretations for the intended use(s).

The excerpts from ETS (2014) and the *Standards* (2014) elucidate two issues which are being addressed in contemporary validity and fairness research. First, both the definitions from ETS and the *Standards* fail to differentiate between the intended and actual uses of test scores. The focus is on fairness as it relates to the intended uses for individual test takers. Moss contends that in a modern era of testing, one which is dominated by accountability systems, “a shift in focus from intended interpretations and uses to *actual* interpretations and uses is necessary” (2016, emphasis original, p. 236).

Second, there is a limited scope of group. Limiting analyses between test-takers based on their gender, ethnicity, native language, etc. places the focus on the individual characteristics of the test taker. This conclusion is in keeping with Leauneanu and Hubley’s (2017) assertion that validation research is typically disconnected from the contextual influences that shape testing situations. Zumbo and colleagues (2015) and later Chen and Zumbo (2017) address these concerns by proposing an ecological model of test taking.

While seemingly disparate theoretical discussions, ecological modeling and accountability systems stress the notion that student testing experiences and data are multilevel in nature. As such, there is a growing body of research focused on similar methodologies across the two research lines. One group of methodologies that are supported in both sets of literature, are analyses of differential item functioning.

The *Standards* (2014, p. 218) formally *defines differential item functioning (DIF)* as “a statistical indicator of the extent to which different groups of test takers who are at the same ability level have different frequencies of correct responses”. DIF is the statistical term that is used to simply describe the situation in which persons from one group answered an item correctly more often than equally knowledgeable persons from another group (Zumbo, 2007). Thus, DIF studies attempt to quantify construct irrelevant variance by assessing whether item responses differ for examinees with the same ability level but in different groups (Dorans & Holland, 1993). As such, DIF studies are considered a key component in the evaluation of the fairness and validity of educational tests (Zwick, 2012).

However, DIF is not synonymous with item bias. Rather, DIF is a necessary but not sufficient condition for item bias (Clauser & Mazor, 1998). If an item exhibits statistical DIF then it should undergo a judgmental procedure to determine the unintended sources of group differences in item difficulty. If these differences are due to an unintended construct that is irrelevant to the attribute being measured then the item is considered *biased* (Camilli & Shepard, 1994). Thus, DIF is a statistical term while bias is a judgmental term. Lastly, *impact* represents between-group differences in test performance caused by a between-group difference on the construct being measured (Ackerman, 1992). Therefore, impact represents results caused by true differences while DIF represents results caused by differences in construct irrelevant variables.

This chapter will provide a brief historical perspective on accountability systems within the United States before turning back to issues of validity, fairness, and DIF in

Chapter Two. Understanding the high stakes associated with accountability testing as well as its pervasiveness is necessary to understanding the current climate of testing and as a result validity, validation, and the assessment of fairness. As a key component of validity and fairness studies, DIF analyses should ultimately be shaped by our current testing landscape. To ground the discussion in current context, examples of the high stakes decisions made under the banner of accountability in North Carolina will be given.

History of Accountability in the United States

Internationally, accountability systems have arguably been the one most powerful trend in education policy in the last twenty years (Volante, 2007). Others have concurred dubbing the increased focus on accountability in the social sector, the ‘age of accountability’ (Hopmann, 2008). Accountability systems combine numerous metrics, including test scores, to result in an overall index or rating for teachers, schools, districts, and educator preparation programs. The myriad of actual uses of test scores due to accountability systems in the United States requires a thoughtful discussion of fairness and validity.

The advent of the Elementary and Secondary Education Act (ESEA) of 1965 marks federal involvement in test-based accountability. Prior to ESEA, several states had introduced statewide testing programs intended to be used for student guidance and the identification of talent (Mazzeo, 2001). Federal involvement in test-based accountability represents an exponentially increasing shift from student-level accountability to accountability at social-unit levels (e.g. teachers, schools, administrators).

Since the 1970s, the United States has increasingly gravitated toward test-based reforms and school regulation. While ESEA was focused on equal opportunity and school improvement, it's reauthorization, the No Child Left Behind Act of 2002 (NCLB), shifts to using standardized tests to hold educators, schools and states accountable to federal guidelines and expectations (Linn, 2006). Although not without dispute, the primary goal of accountability systems is student learning. By holding administrators and teachers accountable, it is hoped that they will be sufficiently motivated to encourage and support student learning. To increase motivation, administrators and teachers are held responsible for student learning (Smith, 2017).

If NCLB represents a shift away from student-level accountability, Race to the Top (RTTT, authorized under the American Recovery and Reinvestment Act of 2009) squarely places accountability at the social-unit level. Specifically, RTTT testing systems are intended to drive teacher effectiveness, school performance, and economic growth through college and career readiness (Deville & Chalhoub-Deville, 2011). To achieve these lofty goals, RTTT mandated many of the design features that testing consortia and independent states must attend to, such as alignment with Common Core State Standards (CCSS).

In 2015, NCLB was succeeded by the Every Student Succeeds Act (ESSA), another reauthorization of ESEA. While ESSA has addressed some of the criticisms lobbied at NCLB, it is still a primarily test based regime, which Mathis and Trujillo (2016, p. 6) characterize as, "a test-driven, top-down, remediate and penalize law." Smith (2017) refers to accountability under NCLB and ESSA as a "punitive testing policy" where formal rewards or sanctions are applied to aggregate scores. ESSA shifts accountability mechanisms to the states, granting states more flexibility. However, accountability systems are still federally approved through the peer review process. Though Adequately Yearly Targets no longer exist schools are still subject to state-imposed sanctions. Identification of schools in need of improvement is still largely determined by test scores.

Federal policies such as NCLB, RTTT, and ESSA represent reform-driven educational initiatives (Deville & Chalhoub-Deville, 2011) which are often referred to as the global education reform movement (GERM) (Sahlberg, 2011). One component associated with GERM in the context of the United States is a focus on test-based accountability policies for schools. Specifically, school performance and student achievement are tied to the process of accrediting, rewarding and punishing schools and teachers. The features of NCLB, RTTT, and ESSA outlined above are in keeping with this definition of GERM.

Local Context. Policies within North Carolina are provided below to fully articulate how states use test scores to make high stakes decisions regarding teachers, schools, and administrators. North Carolina uses end of grade assessments (EOGs) in grades three through eight for Math, English Language Arts and Reading. In grades five and eight students also take a Science EOG assessment. End of course assessments (EOCs) include Biology, English II and NC Math I that are typically taken during the high school years.

The following information is based on legal documentation and reports issued from the North Carolina Department of Public Instruction (NCDPI) for the 2017-2018 academic year unless otherwise stated.

Creating an Accountability Index. The accountability indexes used within North Carolina, as mandated by ESSA, depend on multiple factors. As of 2013, three indexes are provided for each public school: achievement, growth, and performance.

School Achievement. There are five indicators used for calculating the school achievement score. Specifically, (1) the percent of students at an achievement level three or higher (of five) on all of the applicable EOG and EOC assessments, (2) the percentage of graduates who complete NC Math 3 with a passing grade, (3) the percentage of grade 11 students who achieve a 17 or higher on the ACT College Readiness Assessment, (4) the percentage of graduates identified as Career and Technical Education concentrators who meet the Silver Certificate or higher on the ACT WorkKeys Assessment, and (5) the percentage of students who graduate high school within four years (NCDPI, 2017a).

Growth. North Carolina uses a value-added model called the Education Value-Added Assessment System (SAS EVAAS) to assess growth on all EOGs and EOCs taken by students. The growth score results in one of three designations: (1) exceeds expected growth, (2) meets expected growth, or (3) does not meet expected growth (NCDPI, 2017a). Value-added models (VAMs) not only identify growth but attempt to associate growth with particular educators or schools (Castellano & Ho, 2013). As an example, VAM estimates can be interpreted as the average amount of achievement growth an individual teacher contributes to his or her students (Guarino, Reckase, Stacy & Wooldridge, 2015).

Performance. School Performance Grades are a composite of student achievement (80%) and growth (20%) (NCDPI, 2017a). The performance grades are awarded on an A-F scale. Documentation regarding palpable differences between scores on the scale is lacking.

***High Stakes Decisions for Teachers*.** With the passage of the Excellent Schools Act of 2013, teachers are evaluated on six core standards: (1) leadership, (2) establishment of a respectful environment, (3) content area expertise, (4) facilitation of learning, (5) reflection on practice, and (6) contribution to the academic success of students (State Board of Education, 2015). The sixth standard is determined by three-year rolling averages of student growth data observed at the individual (70%) and school level (30%). Teachers who are rated as “developing” or below on any standard or who do not meet expected growth are considered “in need of improvement” and are placed on growth plans monitored by their school administrator (State Board of Education, 2015).

North Carolina ascribes to a merit pay plan which allots bonuses to teachers who perform well. Plans vary by county, but all counties included a measure of growth either through the teacher evaluations (standard six) or through explicit inclusion of a growth indicator. The merit-based plan for Avery County, NC is given as an example. Teachers are awarded points based on five criteria: (1) absences, (2) tenure, (3) school growth, (4) the School Performance Grade, and (5) the number of sub-groups taught (State Board of Education, 2017). Teachers working at schools who received a D or F on the School

Performance Grade scale receive zero points towards their bonuses (more points result in higher bonuses). In this plan, growth is considered in both criteria three and four. In addition to the criteria listed above, at the high school level teacher performance is determined by the number of students who earn industry certifications or credentials or scores on Advanced Placement, International Baccalaureate and Cambridge Advanced International Certificate of Education exams. Teacher bonuses can range from $25 to $6,400 (Helms, 2018).

Merit pay bonuses are therefore dependent upon test scores in numerous ways: (1) through the student and school growth accounted for in teacher evaluations, (2) directly through student growth metrics, (3) through the School Performance Grades, and (4) through additional standardized tests at the high school level.

High Stakes Decisions for Administrators. Principal salaries for the 2017-2018 academic year were determined by two factors: (1) the average daily membership in the school and (2) the accountability growth score for the schools supervised by the principal in two of the last three years (NCDPI, 2017b). Of note, principal salary is not determined by education level, or tenure. For a principal at a small school (up to 400 students) the difference in monthly salary between failing to meet your growth model target and exceeding expectations was $1,030 (NCDPI, 2017c). At a large school (over 1,300 students) the salary difference increases to $1,235 monthly. The difference in salary for a principal in the smallest tier school versus the largest was also $1,030, indicating that performance on the growth model is at minimum given equal weighting as school size.

Similar to teachers, principals and assistant principals are eligible for merit bonuses that range from $1,000 to $15,000 depending on student growth (Helms, 2018).

High Stakes Decisions for Schools. The School Performance Grades result in each school being awarded a single grade from A-F. All schools are required to post their scores on their website. Schools which receive a D or an F are required to notify parents of their score. Outside of these provisions there are not currently rewards or sanctions. However, North Carolina operates Opportunity Scholarships which allots private school vouchers to parents looking to move their child from a public to private school. In 2018, 6,452 students received vouchers to attend private schools compared to just 1,216 in 2015 (EdChoice, 2018). While not directly linked, displeasure with public schools could motivate some parents to seek other options.

Similarly, there are no rewards given to schools which rank highly on the School Performance Grade scale or incentives given to schools who need to improve. Legally, there is no requirement for state officials to allocate funds or resources to these schools.

Given the high stakes associated with accountability systems it is necessary to understand how validity is conceptualized in our current testing environment, how fairness is assessed within validity, and what if any improvements are being suggested given the shift to social-unit level accountability.

Purpose

Accountability systems require sound methodologies for providing evidence of fairness to support the use of test scores at the teacher, administrator, or school level. Previous research has examined multilevel DIF frameworks as solutions for the nested data structure. While scant research has linked multilevel DIF frameworks to the need for teacher and school level fairness evidence in an accountability context (Li, Qin & Lei, 2017), research has addressed the examination of DIF at social-unit levels (e.g. Chen & Zumbo, 2017; Cheong, 2006; Kamata, Chaimongkol, Genc, & Bilir, 2005).

Further research is needed to compare proposed multilevel DIF frameworks. The objective of this study is to investigate the performance of three multilevel DIF frameworks (multilevel Rasch DIF Model, Kamata, 2001; multilevel Mantel Haenszel, French & Finch, 2013; multilevel SIBTEST, French & Finch 2015) under various conditions. The use of test scores at the teacher and school level is the impetus for this study, therefore, DIF will be investigated at the social-unit level.

While understanding the root causes of DIF is a noble cause, and arguably more important than merely flagging contaminated items, the focus of this study will be on identifying social-unit DIF. This is for two reasons. First, multilevel DIF analyses are in their nacency and few studies have focused on multilevel DIF, particularly while framing the study within accountability. Thus, a foundation for future research is necessary. Second, many of the proposed multilevel DIF frameworks are not suitable for identification of the root causes of DIF but could be easily implemented by practitioners to identify multilevel DIF.

Current Study

The current study aims to compare three multilevel DIF frameworks: (1) the multilevel Mantel-Haenszel (MMH; French & Finch, 2013), (2) the multilevel SIBTEST (French & Finch, 2015), and (3) the three-level Rasch model (Kamata, 2001). The goal in such a comparative study is to determine under which conditions frameworks perform well and provide guidelines to practitioners testing for DIF in nested data. This simulation study adds a systematic investigation of social-unit level DIF to a relatively small body of research.

Similar to French and Finch (2013; 2015) and Wen (2014) the current study examines how factors related to social-unit sample size, magnitude of DIF, and intraclass correlation affect the power and Type I error of multilevel DIF frameworks. However, the current study does so in a comparative setting. Additionally, the current study investigates power and Type I error when impact is present. The power and Type I error rates of the three proposed methods have not been examined under conditions of impact. Lastly, the study assumes a balanced approach to DIF which differs from the dominant approach investigated in the initial studies by French and Finch (2013; 2015).

As an additional consideration this study aims to investigate the accuracy of the DIF effect size measures produced by the three multilevel DIF frameworks. Prior studies have focused on power, Type I error, and ability estimation. However, inference tests of statistical significance are just one component to a well conducted DIF study. Significance results need to be paired with effect size measures to allow for meaningful interpretation of results. While the adjustments made to the Mantel-Haenszel and SIBTEST only address the test of significance it is imperative to understand how the effect sizes produced by these methods are affected by nested data.

Research Questions

1. How do three multilevel DIF detection frameworks, the multilevel Rasch model, the multilevel Mantel-Haenszel, and the multilevel SIBTEST, compare to each other in terms of power and Type I error under various conditions?
2. Within each multilevel DIF detection framework, what factors have the strongest influence on power and Type I error (e.g. number of clusters, intraclass correlation, magnitude of DIF, presence of impact, and equivalency of sample sizes)?
3. How accurate are effect size measures produced by the three multilevel DIF detection frameworks?

Organization of the Study

Chapter Two reviews relevant literature on validity, validation, and multilevel validation before turning to DIF and multilevel DIF. The first section describes theoretical considerations of validity and validation in order to provide a strong rationale for why this study is necessary. The second section introduces DIF methods before presenting multilevel DIF frameworks. Chapter Three outlines the data simulation design, modeling approach, and criteria by which the results are evaluated. Chapter Four displays the results of the study. Finally, Chapter Five provides a general discussion of the results, implications for researchers, study limitations, potential future directions, and draws connections back to validation and fairness in an accountability context.

CHAPTER II

LITERATURE REVIEW

This chapter is organized into two main sections. The first focuses on validity and validation as well modern adaptations which address multilevel data. The discussion will start with a brief overview of validity as presented by prominent researchers and the *Standards* (2014). Then the relationship between fairness and validity will be addressed, specifically through a discussion of the role of DIF in validation. Lastly, multilevel validation frameworks which are more closely aligned with accountability systems will be presented. Their theoretical underpinnings and supporting methodologies will be addressed.

The second section will focus on the technical aspects of DIF studies. What follows in this section is a general discussion of DIF, the types of DIF and some generalities of DIF studies, which will be used to lay a foundation for a discussion of the literature regarding multilevel DIF frameworks. The methodology underlying each multilevel DIF framework will be presented before reviewing relevant findings regarding the frameworks power and Type I error under various conditions. Lastly, the method in which anchor items are selected for DIF analyses is discussed as well as the complications presented by multilevel DIF frameworks.

An Overview of Validity

Messick (1989) defines validity as “an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment” (p. 13). Messick argues for a unitary, construct-grounded approach to validity. It is not a characteristic of a test per se, rather a function of test scores and uses. Nor is it a quantity you either have or don’t, it is an evaluative judgment regarding the extent to which empirical evidence supports interpretations of scores and upholds their uses. Our current conceptualization of validity is grounded in the work of Messick. Indeed, the current edition of the *Standards* (2014) frames validity in terms of Messick stating, “validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests” (p. 11).

Kane (2006, 2013) is the counterpoint to Messick’s (1989) philosophical leanings. Kane’s contribution to validity is significant as he lays out a roadmap for practitioners to follow, thus providing actionable guidance rather than theoretical discussion on validity. Kane’s Interpretive Argument (IA; 2006) provides explicit statements of the inferences and assumptions underlying the interpretation and use of test scores. The IA provides practitioners with organized steps to document their validation process. Later, Kane renamed the IA an IUA (Interpretation/Use Argument; 2013) to underscore the importance of interpretation *and* use in the validation process. While the *Standards* (2014) conceptualizes validity in terms of Messick they also incorporate Kane and his concepts of validation. Specifically, the *Standards* define validation as, “a process of constructing and evaluating arguments for and against the intended interpretation of test scores and their relevance to the proposed use" (2014, p. 11).

As Kane's (2006; 2013) conceptualization of validation is widely accepted by the measurement field to the extent that its theoretical underpinnings are presented in a guide published by the three preeminent research organizations it will be presented as a framework for validation. That is not to imply that there are not competing validation frameworks for the consideration of individual level validation. However, Kane's IA/IUA is instructive for understanding some of the issues surrounding validation work and will serve as a foundation for newer multilevel frameworks.

Interpretative Argument-Interpretation/Use Argument. Kane's IA/IUA provides practitioners and researchers a tractable method for accruing validity evidence to support the interpretation and use of test scores. The IA/IUA hinges upon Toulmin's (1958) model for analyzing arguments. Claims are laid out which must be supported by warrants. However, the warrants are generally self-evident and therefore require backing, e.g. evidence. Challenges to the warrant can and should be made which are ideally refutable via the backing. Validity, then, is evaluated in terms of the clarity, coherence, completeness, plausibility, and appropriateness of the claims made (Kane, 2006; Kane, 2013).

An example of an inference and its corresponding claim, warrant, and backing are provided in Table 1. While validation is a context specific process these are given as generalities to highlight the components of Toulmin's argument model.

Table 1

Example of the Extrapolation Inference

Inference	Claim	Warrant	Challenge	Backing
Extrapolation	The items on the test are representative of tasks within the target domain and real world in general.	The construct assessed by the test accounts for the quality of performance in the domain of interest.	The universe of generalization is sufficiently different from the target domain in some way that extrapolation from the universe score to the target score is not legitimate (Kane, Crooks & Cohen, 1999, p. 11).	Evidence is accrued via "think-aloud" protocols where examinees think through the processes they use during a task. If these processes are consistent with other tasks in the target domain then confidence in the inference is strengthened (Kane, 2006, p. 36). A second evidence source is the correlation between the score on the test and an external criterion measure (Kane, Crooks & Cohen, 1999, p. 10).

Figure 1 represents a visual of the validation process as laid out by Kane, Crooks, and Cohen (1999) with the addition of use which is addressed in Kane (2013) though never formally visualized. The addition of use at the end of the validation process is warranted given Kane's separation of evidence evaluation for interpretation and use (2013, p. 47 & p. 56) and is supported in the literature (Chapelle, Enright & Jamieson, 2010). Considered within the decision rule inference, the inference related to use, are consequences. Namely, Kane advocates for the inclusion of three types of consequences: (1) intended outcomes, (2) adverse impact, and (3) systemic effects.

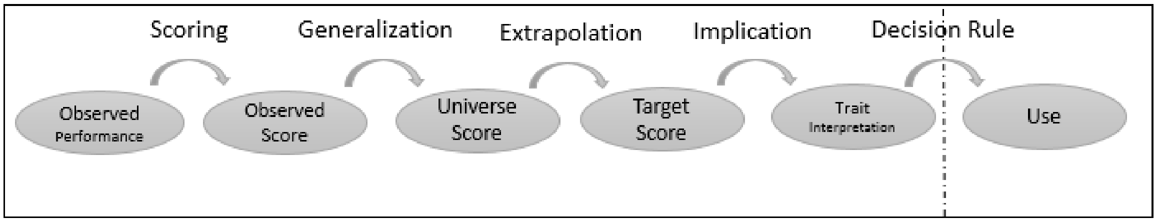

Figure 1. Kane’s Validation Process adapted from Kane, Crooks, and Cohen (1999), Kane (2006), and Kane (2013).

Of note, in Figure 1, while the inferences are presented as separate bridges, a more apt analogy would be as pillars of a bridge. If any one of the inferences which makes up the interpretation argument were to lack strong supporting evidence, then the entire bridge collapses. However, the inferences related to interpretation and use are an exception and remain connected yet separate. Meaning, failure to justify the use of test scores does not automatically invalidate the interpretation of the scores. This separation is represented by the dashed line in Figure 1. Figure 2 represents an expanded view of validation as laid out in Kane (2006) and visualizes terminology introduced in the following discussion.

Examples of the specific claims, evidence, and backing which is needed for each inference will be presented in subsequent sections on validity theory in an accountability context. By situating them as so it is possible to provide a comparison between the claims and evidence needed when considering student level validity and the claims and evidence needed when considering social-unit level validity.

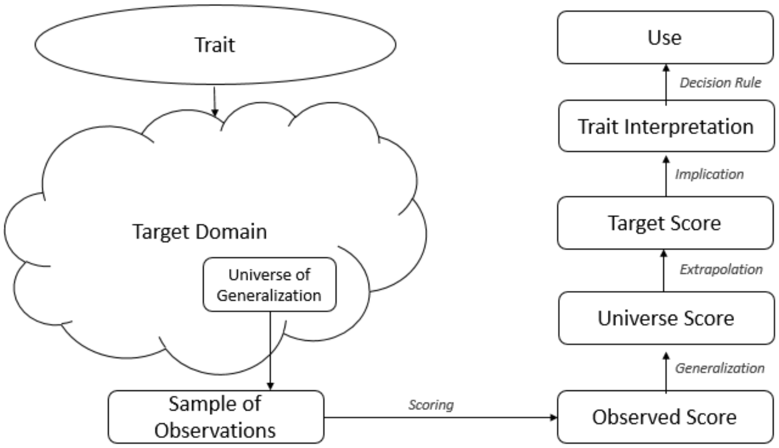

Figure 2. Expanded View of Validation taken from Kane (2006).

The Validation of Use. When considering the validation of use and related consequences there are numerous debates within the field of measurement. First, does the evaluation of use belong in validation and if so is the evaluation of consequences relevant? Second, which uses are we evaluating: intended, actual, or both? Third, if we are to validate the use of a test, then who should be responsible for evaluating it? And lastly, is the evaluation of consequences relevant only to use?

Although Kane (2013) stressed the equal billing that interpretation and use should be given in the validation process, the decision rule inference and consequences are not necessarily accepted as part of the validation process by the measurement field en masse. Researchers span the spectrum from those that believe use has no place in validation (Borsboom & Wijsen, 2015) to those that believe consequences should be evaluated but not under the heading of validity (Cizek, 2012; Lissitz & Samuelsen, 2007) to those who would take Kane’s inclusion of use and consequences farther (Bachman & Palmer, 2010).

Falling on the inclusionary end of the spectrum, Sireci (2016) typifies the case for why use and consequences should be included in validation. He states, “if tests existed only for their scores to be interpreted, but the scores were never used for any purpose, by definition, they would be *usesless tests*. Useless tests have no utility and proposing a definition of validity for them is a fruitless endeavor” (Sireci, 2016, pp. 231-232). As presented by Sireci, interpretation and use are interwoven, and a separation as perceived by Kane (2013) is difficult to achieve.

Hubley and Zumbo (2011) lobby for the inclusion of consequences based on two primary arguments. First, they link the evaluation of consequences to issues of construct underrepresentation or construct-irrelevant variance and the construct. As presented by Hubley and Zumbo, score interpretation and use not only cause consequences but are impacted by their consequences. Particularly convincing is their argument that a consequence of score use may result in revised theories regarding the construct and population of interest.

However, Hubley and Zumbo (2011) argue that not all social consequences are due to sources of invalidity. Specifically, they give the example of financially penalizing schools for children’s poor test performance under NCLB. They view this action as test misuse based on external political beliefs or policies. They contend that such social consequences are outside the realm of validity because they are explicitly linked to test misuse.

However, given that tests are designed explicitly for accountability it is difficult to accept that the current practice of using tests due to political policies represents misuse. Bennett, Kane and Bridgman (2011) specifically reference political policy agendas through their inclusion of intended and unintended negative effects of an assessment system. When listing potential intended effects of two consortium tests used for accountability, the Partnership for Assessment of Readiness for College and Careers (PARCC) and the Smarter Balanced Assessment Consortium (SBAC), they include: (1) making accountability policies better drivers of improvement, and (2) helping education leaders and policymakers make the case for improvement and for sustaining education reforms (Bennett et al., 2011). Both effects are squarely situated in policy, yet Bennett and colleagues also advocate for the involvement of measurement professionals in their evaluation. Clearly within the segment of the field that supports the consideration of consequences within validation disagreement persists as to which consequences should be evaluated.

Hubley and Zumbo's (2011) second argument is that many ***actual*** test uses for measures fall well outside intended test uses and are driven by the desire to bring about personal and social change. Therefore, it is critical to consider the consequences and side effects of measurement in the validation process itself. Some of the specific uses of measures laid out by Hubley and Zumbo are for ranking, intervention, feedback, decision-making, and policy purpose. These uses elucidate a second debate over the inclusion of uses in validation: which uses should we consider, intended, actual, or both?

Sireci (2016) and Moss (2016) each contend that to worry ourselves with intended test uses is not enough. We must also consider the actual uses of tests. According to Moss, actual interpretations and uses are invariably shaped by local users' purposes and depend on the local capacity to use the provided information well. As such, the actual interpretations and uses are far more varied than one might assume (Coburn & Turner, 2012; Moss, 2007). Moss asserts that these local interpretations and uses are ultimately a local responsibility, however, she states that measurement professionals should support local educators and administrators in their validation endeavors.

Within Moss's (2016) discussion of actual test interpretations and uses she makes a direct connection to accountability systems. She states that indirect test uses include score-based incentives intended to raise test scores and the use of test scores to improve schooling. However, the actual uses are far more widespread. In a study by Coburn, Toure, and Yamashita (2009) they found that test scores were used to shape numerous decisions, including: decisions about curriculum adoptions, professional development, and compensation among others. Their work highlights the widespread reach of test results and the multifaceted nature of actual test use.

Similar to Moss (2016), Chalhoub-Deville (2009) addresses the burden of evaluating use through her 'Zone of Negotiated Responsibility', depicted in Figure 3, which offers a sliding scale for determining if the burden of responsibility falls on the test user or developer. The breadth of construct, test use, and time shape the responsibility considerations for test developers and users. As outlined by Chalhoub-Deville, the broader the definition of construct the more the burden of responsibility falls on the test developer. However, as the actual uses diverge from the intended use the responsibility shifts to the test users. The shift in burden related to time is also related to actual uses. As time passes and unintended interpretations and uses of a test persist the test developer can no longer ignore those interpretations and uses and must implement validation research to support them.

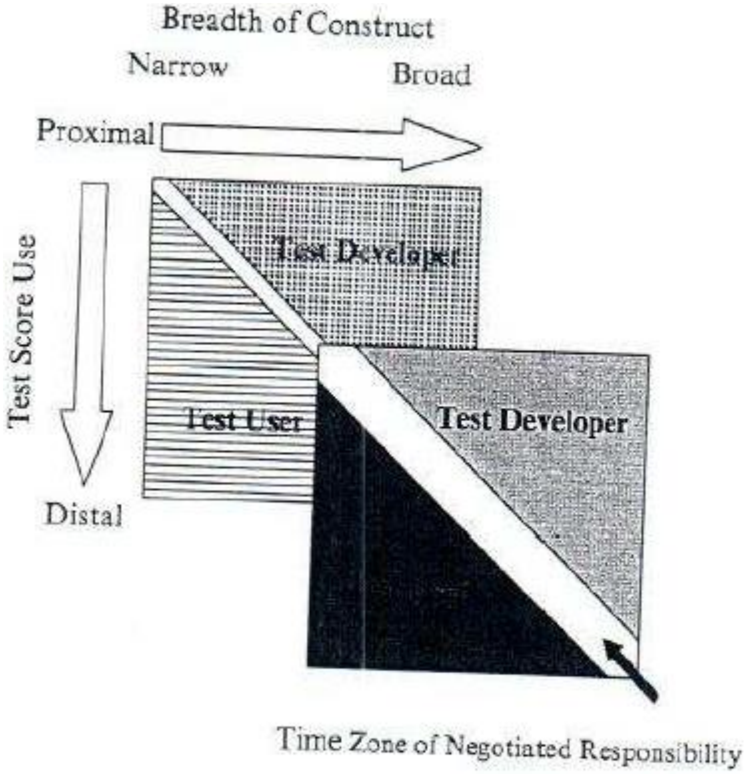

Figure 3. The Burden of Validation Responsibility as Presented by Chalhoub-Deville (2009).

There is not only debate within the measurement field regarding whether consequences should be evaluated at all but also whether consequences are only relevant to score use. Kane (2013) explicated three types of consequences which should be evaluated to support the use inference. Zumbo and Hubley (2016), however, disagree that consequences are only related to the use inference and not to test score inferences or meaning. They argue that researchers have typically linked the evaluation of consequences to use by focusing on the consequences of test misuse, which Zumbo and

Hubley view as outside the realm of validation work. Instead they consider consequences to be the impact or effects of legitimate test score interpretation and use. Therefore, consequences are relevant to the validation process as they are inextricably linked to the meaning of scores.

Validation and Differential Item Functioning

The *Standards* (2014, p. 218) formally defines DIF as “a statistical indicator of the extent to which different groups of test takers who are at the same ability level have different frequencies of correct responses”. Thus, DIF studies attempt to quantify construct irrelevant variance by assessing whether item responses differ for examinees with the same ability level but in different groups (Dorans & Holland, 1993). Within the *Standards* (2014), DIF appears in both the chapter on fairness and validity.

Within the *Standards* (2014), the chapter on validity presents five sources of validity evidence. Specifically, evidence should be collected which relates to (1) test content, (2) response processes, (3) internal structure, (4) relations to other variables, and (5) consequences of testing. DIF studies are proposed as providing validity evidence based on internal structure. DIF studies speak to internal structure by highlighting whether particular items function differently for “identifiable subgroups of test takers (e.g., racial/ethnic or gender subgroups)” (*Standards,* 2014, p. 16). However, the views espoused by the *Standards* are not universally accepted and research has been put forth which challenges our conceptualization of these topics and pushes the field of measurement forward.

The *Standards* (2014) defines construct underrepresentation as “the extent to which a test fails to capture important aspects of the construct domain that the test is intended to measure,” conversely construct-irrelevant variance is “variance in test-taker scores that is attributable to extraneous factors that distort the meaning of the scores” (p. 217). DIF studies allow for the statistical detection of construct-irrelevant variance that is contingent upon group differences. Thus, ensuring that scores obtained from tests are unbiased and reflect the same construct for all examinees (Walker, 2011).

However, Gomez-Benito and colleagues (2018) argue that DIF studies provide broader validity evidence than is implied by the *Standards* (2014). Rather than being limited to evidence for internal structure, they view DIF studies as speaking to the intended interpretation of test scores holistically. Gomez-Benito and colleagues assert that distinguishing DIF from actual differences in the abilities of test takers and determining whether DIF items are measuring the intended construct are fundamental validity issues undertaken in the pursuit of fairness.

Kane’s IA/IUA (2006; 2013) is not the only approach to validation work, indeed there are many approaches presented in the literature. Salient here is Sireci’s deconstructed approach to validation (2016) which relies on explicitly stating the purposes of testing and using the five sources of evidence (as presented in the *Standards, 2014)* to support those explicit purposes. Gomez-Benito and colleagues give explicit examples of how DIF related work can extend to all sources of evidence, thus firmly ingraining DIF analyses in overall validation work. Their examples are presented in Table 2.

Table 2

DIF as a Source of Validity Evidence (Gomez-Benito et al., 2018)

Source of validity evidence	DIF validation work
Test content	Is construct representation similar for identifiable groups of the intended population? Are there differences in the accessibility of test content? Is any content in the items flagged for DIF irrelevant to the construct measured?
Response processes	Do the items tap the same intended process delineated in the test specification for identifiable groups?
Internal structure	Are the relationships among items or part of the test similar for different groups of test takers, i.e. dimensionality? Does an item measure a construct-irrelevant dimension for some examinees?
Relations to other variables	Are the relationships between item/test responses and external criterion following the same pattern for identifiable groups of the intended population?
Consequences of testing	Are unintended consequences of testing arising from construct-irrelevant components or construct underrepresentation? Does the presence of DIF items lead to different pass rates for identifiable groups?

Gomez-Benito and colleagues (2018) make a clear argument for considering DIF analyses as an integral part of validation work. Additionally, they demonstrate how DIF analyses and subsequent analyses related to findings from DIF studies can provide evidence for the other evidentiary components laid out in the *Standards* (2014). In doing so they call for an expansion of what we consider DIF work well past statistical studies to alert measurement professionals to potentially troublesome items.

Zumbo (2009) also views DIF analyses, their methodologies, and impacts as evolving. He categorizes DIF analyses as belonging to one of three generations. The first generation of DIF, more commonly used the term item bias and conflated issues of impact and DIF. The second generation is signified by the adoption of the term DIF and impact, as well as the introduction of new statistical DIF methodologies. The focus during the second generation is on detecting items and distinguishing impact from bias rather than on discerning root causes for DIF. During the second generation of DIF popular methodologies included contingency tables, regression models, and the use of unidimensional and multidimensional item response theory.

The third generation is characterized by what Zumbo (2009, p. 229) refers to as a subtle but extremely important shift to conceiving DIF as occurring, “because of some characteristic of the test item *and/or testing situation* that is not relevant to the underlying ability of interest” (emphasis original). Zumbo’s assertion is salient as we consider the examination of DIF in the era of accountability. This third generation of DIF analyses should consider contextual variables such as classroom size, socioeconomic status, teaching practices, and parental styles. All of which, Zumbo asserts, have typically been ignored in DIF analyses. Including these contextual variables aligns well with the notion that we need to investigate those contextual variables to ensure that the use of test scores at higher social-units levels is valid.

Building on Zumbo’s (2009) description of the third generation of DIF analyses, Gomez-Benito and colleagues (2018) challenge the notion that item bias and DIF remain separate concepts. They assert that psychometricians and measurement professionals have moved past, “statistical analysis which flagged items for DIF to combining statistical findings with substantive explanation regarding the construct underrepresentation and/or construct-irrelevant cause of the differential item performance” (p. 5). This evolution has come about precisely because of advances in modeling (Chen & Zumbo, 2017; Zumbo, Liu, Shear, Olvera, Ark & Ark., 2015).

Multilevel Validation: Context, Accountability, and Methodologies

While validity theory has evolved during the time of NCLB, RTTT, and now ESSA, there remains debate as to if validity theory, as presented above, adequately addresses issues that arise under accountability systems. Chalhoub-Deville (2016) contends that our traditional validity frameworks are inadequate for guiding validation work due to the broad claims at the individual, group, and social system level inherent in GERM movement assessments.

However, there lacks a clear path in the literature for what adequate validity theory looks like under an accountability system. Attempts have been made to outline validation for accountability purposes by modifying existing frameworks (Haertl, 2013), however, these attempts stop short of purposing a new framework explicitly designed for score interpretation and use in the context of accountability. Additionally, systems of item response have been purposed which highlight the multilevel nature of current testing contexts (Chen & Zumbo, 2018; Zumbo et al., 2015). These multilevel theories highlight not only the need for addressing validation for social-unit level test use but also for addressing context as it relates to individual score use and interpretation and for determining root causes of DIF.

Modifying an Existing Validation Framework for Accountability. An example of modifications which can be made to Kane's IA/IUA (2006; 2013) will be given here as it elucidates key issues in using test scores at the social-unit level. Then, a proposed visual framework regarding the process flow of this modified validation framework will be presented.

In a discussion on value-added modeling (VAM), Haertl (2013) highlights the complexity of validation in an accountability context. Haertl adjusts Kane's Interpretative Argument-Interpretation/Use Argument (IA, 2006; IUA, 2013) to discuss the validity of score interpretation and use at the teacher level. However, his discussion and modifications are applicable to other uses of aggregated scores.

Each inference will be presented as it originally was in Kane's (2006; 2013) work and as presented by Haertl (2013) in the context of VAM scores for teacher rewards or sanctions.

Scoring. Traditionally the observed performance would be a student's responses to items, free responses, or performance on a performance assessment. In Kane's IA/IUA (2006; 2013) the observed performance should be understood in this manner. Scores should be comparable across tasks, raters, and test forms (Kane, 2006).

On the other hand, when validating the use and interpretation of VAMs, Haertl (2013) contends the observed performance is that of the teacher. Rather than test items, the scores of examinees are used as measures of their performance. Moving from the teacher's classroom performance to their VAM score (e.g. the scoring inference) must be relatively undistorted by irrelevant factors (Haertl, 2013). Therefore, scores must be free from any systematic bias (Haertl, 2013).

Generalization. Generalization extends the observed performance on a set of tasks to an expected score for a universe of performances that would be considered exchangeable with the current task. The universe of performances is referred to as the universe of generalization and the expected score is referred to as the universe score. At the student-level, generalizability and reliability studies provide estimates of standard errors of measurement and therefore put limits on the precision of estimates of the universe score (Kane, 2006).

When using aggregated scores in a VAM, reliability is still the primary concern of the generalization inference. The stability of VAM scores can be quantified by correlating VAM scores at two points in time or from two sections of the same class (Haertl, 2013). High correlation coefficients would imply stable VAM scores.

***Extrapolation*.** Extrapolation extends the inference from the universe score to the larger universe of performance that is of interest, the target domain. The universe of generalization represents a small subset of the target domain. A key threat to extrapolation is construct-irrelevant variance. Construct-irrelevant variance may result in scores that are systematically higher or lower for identifiable groups of examinees and in inappropriate score interpretations and uses (*Standards*, 2014). Items which introduce construct-irrelevant variance decrease the degree to which we can claim to have supported the interpretation of scores across multiple groups of test takers. According to Kane (2006), extrapolation can be supported via analytic and empirical evidence.

At the student-level, analytic evidence may include: (1) examining the relationship between the processes employed in responding to test tasks and other tasks in the target domain via think alouds and (2) face validity. Empirical evidence may include: (1) criterion-related validity, establishing a direct link between test scores and a valid criterion measure; (2) generalization across new situations or populations of examinees; (3) convergent validity evidence as measured by the correlations between different measures of the trait; and (4) results from a multi-trait multi-measurement study.

Use of aggregate scores at the social-unit level, such as in a VAM, requires evidence linking the VAM scores to broader notions of teacher effectiveness to support the extrapolation inference. When adapting Kane's IA/IUA (2006; 2013) for use with teachers and VAMs, the concern remains: is there construct-irrelevant variance which is jeopardizing our score interpretations and uses?

Haertl (2013) contends that evidence must be collected which addresses the following questions: (1) how well does the VAM score correspond to other kinds of information about teaching quality?, (2) how much do the estimates change if a different test is used?, (3) do the achievement tests reflect the range of desired cognitive outcomes, and (4) is it possible to extrapolate beyond test scores to a broader range of schooling outcomes?. This evidence may be collected through qualitative methods such as classroom observations and in-depth interviews. Empirical evidence includes calculating VAM scores using different subtests from a given assessment for the same students and teachers.

Implication. Implication extends the target score to a verbal description of the trait. Accountability legislation impacts the implication inference for students and at the social-unit level. Specifically, NCLB mandated students be categorized into three proficiency levels (basic, proficient, advanced) based on test scores. Some states added additional levels related to college and career readiness (e.g. North Carolina). Evidence which can support the student level interpretation includes: (1) longitudinal studies of school and college performance and (2) correlational studies between school performance (e.g. GPA) and test scores to support labels such as "grade level proficient" (Beimers, Way, McClarty, & Miles, 2012; O'Malley, Keng, Miles, 2012).

While Haertl (2013) does not provide specific evidence needed to support the implication inference, evidence accrued through qualitative methods documenting teacher quality and efficacy would support the use of VAM scores to make categorizations of teacher efficacy.

Use. Use, as described by Kane (2013), specifically regards evaluating three types of consequences: (1) intended outcomes, (2) adverse impact, and (3) systemic effects. The extent to which consequences must be evaluated and supported with evidence is tied to the stakes of the testing program. According to Kane, the evaluation of consequences is imperative if decisions to be made are high stakes.

While sanctions at the student level are not currently a focus of accountability policy, the emphasis on the use of test scores as a vehicle for student learning and achievement has student level implications. Even when test scores are explicitly designed for use at the social-unit level, the resulting policy decisions would impact students, teachers, and the larger community. Forer and Zumbo (2011) refer to consequences which are felt at multiple levels as cross-level consequences.

The same types of consequences are relevant at the social-unit level. Haertl (2013) describes unintended negative consequences of the use of VAM scores to make high stakes decisions as: (1) a decrease in career satisfaction as teaching becomes more prescriptive due to pressures to teach to the test, (2) more competition and less cooperation among teachers, (3) less supportive peer and mentoring relationships with new teachers, and (4) resentment or avoidance of students who do not learn easily.

The discussion of consequences here is limited to the ***use*** of student level or aggregate test scores.

Visualizing a Framework for Accountability Systems. Haertl (2013) provides an example of modifying Kane's IA/IUA. However, in this framework there lacks an interaction between student and social-unit level evidence accumulation. Within an accountability system, validation would be undertaken at both levels. Figure 4 outlines one potential framework for the organization of collecting student- and social-unit level evidence.[1]

The relationship between the student- and social-unit level is comparable to the relationship Kane (2013) lays out for interpretation and use. A lack of evidence to support social-unit level validation does not necessarily invalidate the interpretations and uses of test scores at the student level. Nor does support for validation at the social-unit

[1] Figure 4 represents a visualization based on work presented by Karen Hoeve, Jeremy Acree, and JB Weir in UNCG ERM's Validity and Validation course in Fall 2017.

level validate the interpretation and use of test scores at the student level. However, a lack of evidence at the student-level would be a red flag that aggregating scores is inadvisable. Additionally, some analyses may provide relevant evidence for both levels.

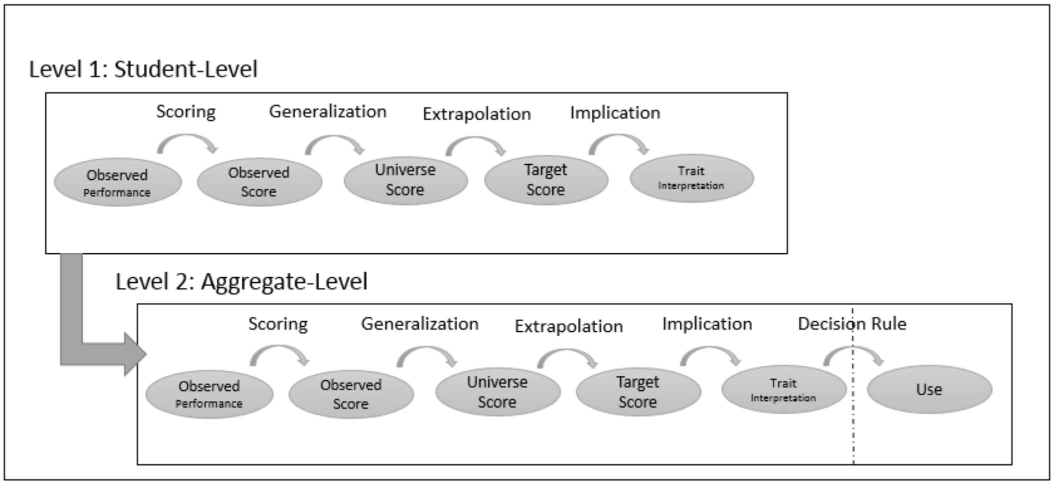

Figure 4. Kane’s IUA Adapted to Reflect Test Use in an Accountability System.

The treatment of use in this framework is variable and is dependent upon the intended and actual uses of a test and accountability index. For instance, current reform-driven educational initiatives link school performance and student achievement to the process of accrediting, rewarding and punishing schools and teachers (Chalhoub-Deville, 2016). Under such systems the intended test use is not at the student level rather at the social-unit level. While the intended use would need to be evaluated at the social-unit level, it is likely that there would be actual test uses which spanned both the student and social-unit level. Additionally, there may be unintended negative effects which span both levels. Lastly, improved student learning and performance is at the heart of accountability systems and while the consequences are typically at the social-unit level they are enacted to bring change about at the student level. Thus highlighting, the necessity of linking student and social-unit level validation.

An Alternative Conceptualization of Multilevel Validation: Multilevel Constructs and Contextualized Response Processes. Zumbo and Forer (2011) address validation by considering multilevel constructs. They define multilevel constructs as those which have meaningful uses and interpretations at the individual and social-unit levels. To interpret or use data at the aggregate (e.g. social-unit) level, one must also present validity evidence at the aggregate level (Zumbo & Forer, 2011).

Zumbo and Forer (2011) present a step by step procedure for conducting multilevel construct validation, which is adapted from Chen, Mathie, and Bliese (2004). The five steps in their framework are:

1. Define the construct across levels of analysis;
2. Articulate the nature of the aggregate construct;
3. Determine the psychometric properties of the construct across levels of analysis;
4. Ensure that there is construct variability across levels of analysis;
5. Examine the function of the construct across levels of analysis.

Forer and Zumbo (2011) utilize these steps to provide construct validation evidence for the Early Development Instrument, a school readiness assessment of kindergartners. While the test is an individual assessment, it is purely designed to be used for policy and planning by educators and administrators. Thus, it represents a multilevel construct. Their utilization of the multilevel construct validation process results in two significant conclusions which are relevant to multilevel validation in general. First, multilevel construct analysis is necessary to ensure that the structure and function of a multilevel measure are isomorphic across any level of aggregation. Second, one cannot assume the construct is the same at all levels of aggregation merely because they represent aggregation. Significantly, dimensionality must be addressed at all plausible levels of aggregation.

Multilevel validation should not only address issues of multilevel constructs but also the contextual nature of testing. Zumbo and colleagues (2015) and later Chen and Zumbo (2017) present an ecological model for considering response processes. Response processes are defined as mechanisms that generate observed test score variation (Embretson, 2010; Messick, 1995). Zumbo and colleagues (2015) consider evidence focused on why and how people respond to items as they do to be central evidence for measurement validation. Within their ecological framework, they explicitly address multilevel DIF analyses as integral to providing validation evidence.

Zumbo and colleague's (2015) work utilizes Brofenbrenner's (1994) ecological systems theory which is popular in the social sciences as a basis for understanding the interaction between examinees and test items. They also build upon Chalhoub-Deville's (2003) ability-in language user-in context theorem while acknowledging that her description of construct is generalizable beyond language assessment. Particularly salient for the ecological model of item responses is the notion that "ability and context features are intricately connected and it is difficult or impossible to disentangle them" (Chalhoub-

Deville, 2003, p. 372). Due to the interconnection of contextual features, ability, and item response a broader framework for examining response processes is necessary.

Such a framework is presented in the ecological model of item responding (Chen & Zumbo, 2017; Zumbo et al., 2015). Zumbo and colleagues explicitly view this ecological model as the foundation for the statistical and psychometric methodology of DIF analysis.   This model provides a contextualized and embedded view of response processes and is presented in Figure 5. In particular this model is appropriate for educational testing, but adjustments could easily be made to better suit licensure and certification settings.

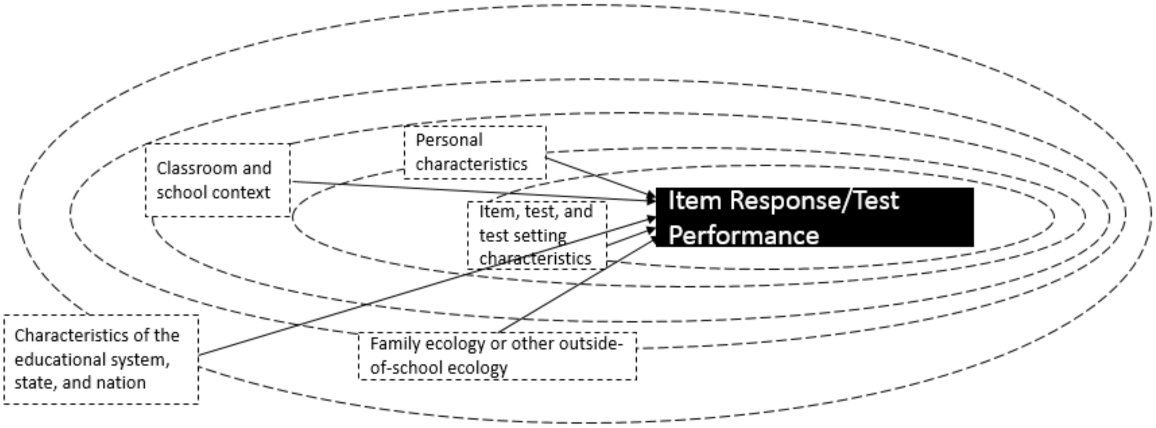
Personal characteristics
Classroom and school context
Item Response/Test Performance
Item, test, and test setting characteristics
Characteristics of the educational system, state, and nation
Family ecology or other outside-of-school ecology

Figure 5. Zumbo et al.'s (2015) Ecological Model of Test Taking.

At the most immediate level are item and test properties, such as the content of the test, format of the test, and the test's psychometric properties (Chen & Zumbo, 2017; Zumbo et al., 2015). The next layer, which is typically the focus of DIF studies, includes student characteristics. However, it is the outer levels that are particularly informative in an accountability context. Moving outward the next layers include the classroom and school context, family or outside of school ecology, and the characteristics of the education system and nation state.

Framing validation work within an ecological framework has multiple significant outcomes. First, understanding how social-unit level characteristics impact item response and test performance becomes salient. This outcome is particularly relevant in an accountability context where score use is happening at the social-unit level.

Second, moving to an ecological framework challenges the notion that DIF analyses be limited to identifiable subgroups as proposed by the *Standards* (2014). Zumbo and colleagues (2015) argue for the use of latent class modeling to detect DIF between groups which are not identifiable via observed characteristics. Their reasoning hinges on the fact that levels of the ecological system may interact in ways that are unobserved. For instance, as opposed to focusing on gender based DIF analyses, a more modern take on DIF would move past focusing on biological sex differences and consider gender as a social construct. Therefore, considering the influence of institutionalized gender roles, classroom size, socioeconomic status, parental styles and how all these factors may shape the construct of gender (Zumbo et al., 2015).

The study of DIF for observable characteristics beyond the student level represents a middle ground to DIF analyses as they are currently conducted and the latent approach espoused by Zumbo and colleagues (2015). To extend the gender example given above, perhaps student level differences due to gender vary in severity dependent upon a social-unit variable, such as the gender of a teacher. Numerous interactional factors could exist such as the average socio-economic status of the school, school resources, the proportion of English Language Learners (ELLs) in the school, etc.

Situating Validation within the Ecological Model. As presented by Zumbo and colleagues (2015) and Chen and Zumbo (2017), an ecological stance can be taken for the modeling of response processes. However, the ecological model could reasonably serve as a broader framework through which the entire validation process can be viewed. It stands to reason that as new methodologies are introduced and we are faced with increasing demands for strong validation evidence in the face of expanded test uses that our frameworks will evolve and broaden as well.

Figure 6 presents all of validation as existing within the large ecological model. Validation work replaces item response processes to signify that all elements of the validation work exist within the ecological model, not just work related to response processes. This view is closely aligned with the adapted accountability validation framework. However, it builds upon that work by stressing that the individual examinee is situated within the larger contextual levels. Thus, validation requires a weighing of evidence at all applicable levels.

If the multilevel adaptation of an Interpretation/Use argument fits squarely in *The Third Generation of DIF*, then an ecological approach to validation advances validation even further. While the two-pronged approach addresses the validation of interpretation and use at the social-unit level, an ecological validation model addresses the interaction of item, examinee, and societal characteristics. The ecological model would address the same inferences as the multilevel IUA while necessitating additional inferences to describe those interactions. It is beyond the scope of this paper to propose those new inferences and exhaustively list the methodologies which would be used to provide evidence to support them. However, future research into such frameworks is warranted.

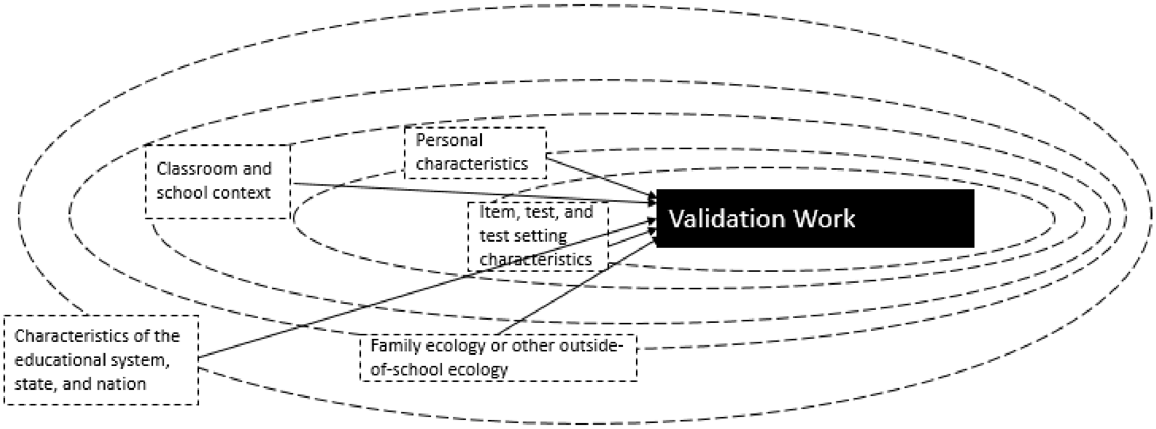
Personal characteristics
Classroom and school context
Validation Work
Item, test, and test setting characteristics
Characteristics of the educational system, state, and nation
Family ecology or other outside-of-school ecology

Figure 6. An Ecological Model of Validation.

Supporting Methodologies for Modified Validation Frameworks. Even if modifications to existing frameworks, such as those documented above, are widely adopted, a lingering question remains: do current methodologies provide sufficient evidence to support the use of test scores to make decisions regarding teachers, schools, and administrators?

Haertl (2013) outlines some applicable methodologies, albeit in a very specific context, through a discussion of current studies. However, he glosses over the methodologies used to support the implication and use inference. Additionally, our current practices do not necessarily reflect best practices.

Zumbo and Forer (2011), Zumbo et al. (2015), and Chen and Zumbo (2017) introduce multilevel DIF analyses as a component to collecting evidence regarding the construct and response processes. Modern DIF analyses, as characterized in Zumbo's (2009) *Third Generation of DIF*, encourage new ways of conceptualizing DIF and fit nicely within multilevel frameworks. Additionally, some multilevel DIF analyses address issues raised by Gomez-Benito and colleagues (2018) and Zumbo (2009) regarding a shift to understanding the root causes of DIF. Multilevel DIF analyses have the potential to further our current practices by addressing score interpretation and use at the social-unit level while offering a more nuanced understanding of the root causes of DIF.

The application of DIF analyses to support the interpretation and use of test scores at the social-unit level raises two issues. First, DIF analyses are primarily conducted with covariates at the student not social-unit level. Second, these analyses typically ignore the naturally nested structure of educational data.

Regarding the first issue, testing programs historically advocate for investigating DIF for groups that experience or research has indicated are likely to be adversely impacted by construct-irrelevant influences on their test performance (ETS, 2014). Typically, analyses include groups which have been discriminated against on the basis of ethnicity, disability status, gender, native language, or race and are defined legally. The *Standards* (2014) gives racial, ethnic, and gender subgroups as examples when discussing DIF analyses. However, none of these analyses provide sufficient evidence to support the interpretation and use of test scores at the social-unit level.

A local example will be given to highlight the potential issues missed by merely focusing DIF on individual characteristics. Within North Carolina, School Performance Grades are reported on an A-F scale, with an emphasis placed on schools receiving a C or higher in a variety of publicly available documentation provided by NCDPI. When School Performance Grades are broken down by socioeconomic status, 35% of economically disadvantaged schools (schools reporting 50% or higher student poverty rates) failed to meet the criteria for a grade of C or higher in comparison to only 4% of non-economically disadvantaged schools (NCDPI, 2017d). Put another way, economically disadvantaged schools accounted for 91.8% of the D ratings in North Carolina and 98% of the F ratings.

While EOG and EOC test scores are only a single component, standardized assessments make up four of the five indicators which go into the School Performance Grades. Conducting DIF analyses at the student level would highlight statistically significant differences due to students' status as economically disadvantaged or not when matched on ability. However, it does not account for systemic differences which may exist at the school level due to issues caused by pervasive poverty. Since the School Performance Grades are used at the school level such an analysis could provide illuminating information regarding the comparability of scores between schools.

Regarding the second issue raised by conducting DIF analyses to provide social-unit level data, students are nested within teachers, schools and districts. Ignoring the nested structure of the data has been shown to result in substantial inflation of the Type I error rate under certain conditions in DIF analyses (French & Finch, 2010). Traditional DIF analyses fail to account for the hierarchical structure of student data.

Differential Item Functioning

Technical aspects of DIF analyses will now be presented. While the previous section contends that traditional DIF analyses are inadequate to address Zumbo's *Third Generation of DIF* and are in many cases statistically inappropriate, this section will proceed with laying a foundation focused on traditional DIF analyses. In order to understand the future of these studies, it is necessary to have a historical understanding. A discussion of more advanced multilevel DIF frameworks will follow.

Approaches for DIF detection can be quantified as matching on ability level through either the observed score (e.g. total test score) or latent variable (e.g. theta) (Millsap & Everson, 1993). Both approaches assume unidimensionality of the data but differ on the criteria used for matching. Potenza and Dorans (1995) added an additional classifying feature, whether procedures use a functional form for the relationship between item score and the matching variable (parametric) or do not (non-parametric).

Table 3

Classification of DIF Detection Methods

Type of Matching Variable	Parametric	Nonparametric
Observed Score	LR DIF	Mantel-Haenszel Standardized Mean Difference
Latent Variable	General IRT Limited Information IRT Loglinear IRT IRT-D^2 Lord's χ^2	SIBTEST

Typically, DIF analyses are undertaken between two groups, a reference and focal group. The focal group is the group of interest or the group for which item bias is a concern. However, an item may exhibit DIF or item bias towards the reference group. DIF is treated symmetrically when bias is any invalid difference between groups rather than only negatively impacting minority groups (Zieky, 1993).

In dichotomously scored items three types of DIF exist: (1) uniform, (2) nonuniform, and (3) nonuniform crossing. The simplest is uniform DIF. Uniform DIF exists when the statistical relationship between response on an item and group membership is consistent across all levels of the matching criterion (Mellenbergh, 1982). When utilizing the Rasch model only uniform DIF is possible as the discrimination parameter (slope) is constrained to be 1.0. Nonuniform DIF exists when the relationship is not constant across levels of the matching criterion due to an interaction effect between group and the matching variable (Mellenbergh, 1982). Nonuniform DIF can take on two forms. DIF may be nonuniform but the item characteristic curves (ICCs) never cross or it could result in DIF in which the ICCs cross. Figure 7 graphically presents the three types of dichotomous DIF.

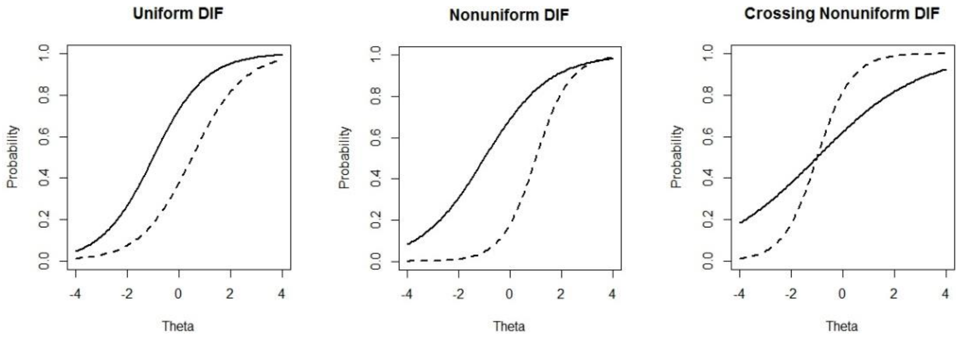

Figure 7. Types of Dichotomous DIF.

Effect Size. Ultimately, there comes a time when a decision must be made regarding each item. Should an item be labeled as bias-free or does it require additional scrutiny of item content and possible removal (Penfield & Camilli, 2007)? As unrealistic as it is to investigate DIF for all possible groups, it is equally unrealistic to extensively investigate bias for all items. Therefore, ETS (2014) recommends investigating differences between groups large enough to have *practical consequences*. However, inferential test statistics are not appropriate measures of the practical size of DIF, and they should not be used as effect sizes (Camilli, 2006). As a result, practitioners have gravitated towards the use of inferential test statistics and measures of effect size (Penfield & Camilli, 2007).

Coupling an inferential test statistic with a measure of effect size draws the distinction between statistical significance and practical significance. Put simply, measures of effect size are a way of determining whether or not the DIF that was detected is sufficiently large to be meaningful (Kim, Cohen, Alagoz, & Kim, 2007). Evaluation of DIF through both a statistical test and measure of effect size reduces false identification rates. In large sample sizes, statistical tests may be significant while the effect size is small (Hidalgo & Lopez-Pina, 2004; Kim et al., 2007).

Traditional DIF Detection Procedures

Throughout the discussion of traditional DIF detection methods the corresponding measures of effect size and classification categories will be presented.

Non-parametric Observed Score Matching Approaches.

Mantel-Haenszel. The Mantel-Haenszel (Holland & Thayer, 1988; Mantel & Haenszel, 1959) is computationally one of the simplest DIF detection methods and therefore particularly popular as evidenced by its use at Educational Testing Service (ETS; Dorans & Holland, 1993).

The Mantel-Haenszel approach is best understood via 2x2x*K* contingency tables (see Table 4). Where *k* represents the levels for the matching variable, *k*=1,2,…*K*. In the case of the Mantel-Haenszel, the matching variable represents the observed total test score for examinees.

Table 4

Contingency Table

	Correct Response (1)	Incorrect Response (0)	Total
Reference Group	A_k	B_k	Nr_k
Focal Group	C_k	D_k	Nf_k
Total	N_{1k}	N_{0k}	N_k

Table 4 consists of the frequencies of correct and incorrect responses for the reference and focal groups respectively. Therefore, A_k is the frequency of correct response in the reference group, B_k is the frequency of incorrect response in the reference group, C_k is the frequency of correct response in the focal group, and D_k is the frequency of incorrect response in the focal group for the *kth* level of the matching variable.

At the k^{th} level, N_{1k} and N_{0k} are the number of examinees who answer the studied item correctly and incorrectly, respectively, N_{rk} and N_{fk} are the number of examinees in the reference group and the focal group, respectively, and N_k is the total number of examinees.

Under the Mantel-Haenszel approach, the null hypothesis to be tested is that the odds of correct response on an item across all levels of the matching variable is the same for the focal group and the reference group (Dorans & Holland, 1993). The null hypothesis can be expressed as:

$$H_0 \colon \frac{A_k}{B_k} = \frac{C_k}{D_k}. \qquad (1)$$

The Mantel-Haenszel chi-square test with one degrees of freedom is associated with the null hypothesis,

$$MH\ CHISQ = \frac{(|\sum_k A_k - \sum_k E(A_k)| - 0.5)^2}{\sum_k Var(A_k)}, \qquad (2)$$

where

$$E(A_k) = \frac{N_{rk} N_{1k}}{N_k}, \qquad (3)$$

and

$$Var(A_k) = \frac{N_{rk} N_{fk} N_{1k} N_{0k}}{N_k^2 (N_k - 1)}. \qquad (4)$$

The 0.5 that is subtracted from the numerator of the MH CHISQ is a continuity correction, designed to improve the approximation of a discrete distribution with a continuous distribution (Zwick, 2012).

Effect Size. The *MH D-DIF* index, developed by Holland and Thayer (1988), leads to a measure of effect size and is calculated as follows:

$$\hat{\alpha}_{MH} = \frac{\sum_k A_k D_k / N_k}{\sum_k B_k C_k / N_k}, \tag{6}$$

and

$$MH\ D - DIF = -2.35 \ln(\hat{\alpha}_{MH}). \tag{7}$$

If the *MH D-DIF* index is smaller than 0, then the item reflects possible bias against the focal group and if the index is larger than 0, then the item reflects possible bias against the reference group. The *MH D-DIF* index is on the ETS delta scale of item difficulty (Zwick, 2012). Thus, a value of -1 means that the item is estimated to be more difficult for the focal group by an average of one delta point, conditional on ability. The statistic can also be conveyed in terms of the odds ratios, *MH D-DIF*=-1 equates to $\hat{\alpha}_{MH} = 1.530$. A value of 1.530 means that the odds of answering the item correctly for the reference group are approximately 50% higher than the odds of answering correctly for comparable members of the focal group.

The significance test and measure of effect size are used to classify items as having A, B+, B-, C+, or C- DIF. Pluses indicate items which favor the focal group, minuses favor the reference group. The classification rules are defined in Table 5.

Table 5

Effect Size Measures for the Mantel-Haenszel

Categorization	Level of DIF	Classification Rule
A	Negligible	*MH CHISQ* is not significant at the 5% level **or** |*MH D-DIF*| < 1.
B	Moderate	If an item does not meet the criteria for a A-item or C-item it is classified as a B-item. *MH CHISQ* > 3.84 **and** |*MH D-DIF*| ≥ 1.
C	Large	|*MH D-DIF*| ≥ 1.5 **and** is significantly greater than 1 in absolute value at the 5% level. |*MH D-DIF*| is significantly greater than 1 if: (|*MH D-DIF*|-1)/*SE*(*MH D-DIF*) > 1.645.

According to personal communication from Neil Dorans as cited in Zwick (2012), C items are to be avoided though B items while undesirable can be tolerated.

SIBTEST. The SIBTEST was developed for the purpose of uniform DIF detection (Shealy & Stout, 1993). The corresponding, CSIBTEST is used for the detection of nonuniform DIF (Li & Stout, 1996). The SIBTEST will be the focus of this section and corresponding multilevel discussions. Although the SIBTEST is meant to be a non-parametric latent variable approach, a common modification is to use the observed total test score as the matching variable.

Items are initially divided into two non-overlapping subsets: (1) a *valid subtest*, which contains items that are assumed to measure the ability the test is designed to measure, and (2) a *suspect subtest*, which contains items being tested for DIF (Bolt, 2000). Scores on the valid subset only are used to match examinees in order to test items from the suspect subtest for DIF. DIF is assessed through estimating the function

$$B(\theta) = P(\theta, R) - P(\theta, F), \tag{8}$$

where $P(\theta, R)$ and $P(\theta, F)$ indicate the probability of a correct response on an item in the reference and focal groups, respectively. The function is integrated over theta to produce the SIBTEST DIF index,

$$\beta_{UNI} = \int B(\theta) f_F(\theta) d\theta, \tag{9}$$

where $f_F(\theta)$ is the density function for theta in the focal group.

β_{UNI} is a weighted expected score difference between reference and focal group examinees of the same ability on the item. Test scores on the valid subtest can be substituted in for theta, producing an estimate of β_{UNI},

$$\hat{\beta}_{UNI} = \sum_{l=0}^{N} p_l (\bar{Y}_{Rl} - \bar{Y}_{Fl}), \tag{10}$$

where

N = the maximum possible valid subtest score,
$\bar{Y}_{Rl}$ = mean scores on the suspect item(s) for reference group examinees having valid subtest score l
$\bar{Y}_{Fl}$ = mean scores on the suspect item(s) for reference group examinees having valid subtest score l, and
p_l = the proportion of focal group examinees obtaining valid subtest score l.

However, should impact exist and the groups have different distributions of theta, the estimate of β_{UNI} will be biased toward indicating DIF favoring the group with higher ability (Bolt, 2000; Shealy & Stout, 1993). Thus, SIBTEST uses a regression correction method to match examinees and compare their performances on studied items. First, SIBTEST computes a regression equation for each group that estimates true score on the valid subtest as a function of valid subtest observed score. Separately for each group, the following equation is calculated:

$$\hat{M}_{gl} = \frac{\bar{Y}_{g,l+1} - \bar{Y}_{g,l-1}}{\hat{V}_g(l+1) - \hat{V}_g(l-1)}, \qquad (11)$$

where

g = group, either focal or reference,
$\bar{Y}_{g,l+1}$ = mean scores on the suspect item(s) for the given group examinees having valid subtest score l+1,
$\bar{Y}_{g,l-1}$ = mean scores on the suspect item(s) for the given group examinees having valid subtest score l-1, and
$\hat{V}_g(l)$ = valid subtest true score estimates for either group.

$\hat{M}_{gl}$ is then used to calculate the adjusted mean score of the suspect item(s) for each group at a valid subtest score l using the following equation:

$$\bar{Y}_{gl}^{*} = \bar{Y}_{gl} + \hat{M}_{gl}\left(\hat{V}(l) - \hat{V}_{g}(l)\right), \tag{12}$$

where

$\hat{V}(l)$ = mean of the valid subtest true score estimates $\hat{V}_R(l)$ and $\hat{V}_F(l)$, and
$\bar{Y}_{gl}^{*}$ = estimated suspect item true score in group g for examinees who have a common estimated valid subtest true score $\hat{V}(l)$.

Second, a revised estimate of $\hat{\beta}_{UNI}$ is calculated,

$$\hat{\beta}_{UNI} = \sum_{l=0}^{N} p_l(\bar{Y}_{Rl}^{*} - \bar{Y}_{Fl}^{*}). \tag{13}$$

As with other DIF statistics, $\hat{\beta}_{UNI}$ has a value of zero under conditions of no DIF. A test statistic for hypothesis testing can be calculated using $\hat{\beta}_{UNI}$,

$$SIB = (\hat{\beta}_{UNI})/\hat{\sigma}(\hat{\beta}_{UNI}), \tag{14}$$

where the sample variance is calculated as,

$$\hat{\sigma}(\hat{\beta}_{UNI}) = \left[\sum_{l=0}^{N} p_l^2\left(\frac{\hat{\sigma}^2(\mathrm{Y}|l,R)}{N_{Rl}} + \frac{\hat{\sigma}^2(\mathrm{Y}|l,F)}{N_{Fl}}\right)\right]^{1/2}. \tag{15}$$

Testing for bi-directional DIF, either against the reference of focal group, the null hypothesis, $H_0\colon \beta_{UNI} = 0$, is rejected at α=.05 if |SIB| exceeds the 97.5% cutoff from the standard normal table.

Effect Size. Roussos and Stout (1996) suggested guidelines for interpreting the value of the SIBTEST DIF index and classifying items, which are derived from the ETS

Delta Scale for the Mantel-Haenszel procedure. These guidelines are presented in Table 6. There is not a fixed mathematical relationship linking $\hat{\beta}_{UNI}$ to *MH D-DIF,* but Shealy and Stout (1993) recommend using -15*$\hat{\beta}_{UNI}$ to achieve a comparable value. This relationship is based on empirical data sets.

Table 6

Effect Size Measures for SIBTEST

Categorization	Level of DIF	Classification Rule
A	Negligible	$\lvert\hat{\beta}_{UNI}\rvert$<0.059
B	Moderate	$0.059 \ \leq \lvert\hat{\beta}_{UNI}\rvert < 0.088$
C	Large	$\lvert\hat{\beta}_{UNI}\rvert \geq 0.088$

Parametric Latent Variable Matching Approaches. In addition to non-parametric DIF detection methods, there are multiple parametric DIF detection approaches based on item response theory (IRT). Prior to a summary of popular IRT DIF detection methods, three common IRT models for dichotomous data will be presented.

IRT models describe the relationship between item characteristics and person latent traits via a probability function. The Rasch Model (Rasch; 1960) can be viewed as an item response model in which the characteristic curve is a one-parameter logistic function. The equation can be written as

$$P_{ij}\left(Y_i = 1 \middle| \theta_j\right) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)}, \qquad (16)$$

where θ_j is an examinee's ability level and $b_i$is the difficulty of the *i*th item, e.g., the point at which an examinee has a 50% chance of correctly responding to an item under the Rasch model (Hambleton & Swaminathan, 1985).

The two-parameter logistic model (2-PL; Birnbaum, 1968) differs from the Rasch model due to the addition of an item-specific discrimination parameter (a_i), which can be thought of as how well an item separates individuals along the ability continuum. It is parameterized as

$$P_{ij}\left(Y_i = 1\middle|\theta_j\right) = \frac{\exp\left(a_i\left(\theta_j - b_i\right)\right)}{1 + \exp\left(a_i\left(\theta_j - b_i\right)\right)}. \tag{17}$$

Lastly, the three-parameter logistic model (3-PL; Birnbaum, 1968) builds upon the 2-PL through the addition of a pseudo-guessing parameter (c_i). The addition of the pseudo-guessing parameter raises the lower asymptote of the ICC and represents the probability of examinees with low ability correctly guessing the answer to an item.

Magis, Beland, Tuerlinkcx, and de Boeck (2010) classify parametric DIF approaches into three main types: (1) the likelihood ratio test method (LRT; Thissen, Steinberg, & Wainer, 1988), (2) Lord's chi-square test (Lord, 1980), and (3) Raju's method (Raju, 1988). Briefly each method will be summarized.

The LRT method (Thissen et al., 1988) consists of fitting two IRT models, a compact model with identical item parameters for each group and a modified model where item parameters are allowed to vary. The significance of these parameters is tested by the likelihood ratio test. Which parameters are allowed to vary is dependent upon the model, i.e. in the Rasch model only the difficulty parameters can vary.

Lord's chi-square test (Lord, 1980) is based on the null hypothesis of equal item parameters in the reference and focal group. This approach has also been extended to account for more than two groups of interest (Kim, Cohen & Park, 1995). It differs from Raju's method (Raju, 1988) which compares the ICCs between the reference and focal group. Raju's method assumes the true area between the two curves is zero. While any IRT model can be considered with Raju's approach, the pseudo-guessing parameter for both groups must be constrained to be equal.

Multilevel DIF Detection Frameworks

Multilevel DIF frameworks offer multiple benefits over their traditional counterparts. Neither statistical significance tests for DIF detection nor effect size measures themselves provide insight into understanding why DIF occurs, an essential exercise for test developers (Kim et al., 2007). Understanding the cause of DIF can assist test developers in determining if DIF truly represents an unfair advantage on an item and aid in the development of fairer tests (Penfield & Camilli, 2007). In addition to the necessity for multilevel DIF detection presented in Chapter I, multilevel DIF frameworks offer a potential solution to determining the root cause of DIF. By modeling DIF at multiple levels, item level characteristics can be introduced (*differential facet functioning*; Cid, 2009) as well as social-unit level characteristics such as opportunity to learn and teacher effectiveness (Burkes, 2009; Wen, 2014). These additional covariates can be used to explain the variability in item parameter estimates between groups of examinees. According to Zumbo and Hubley (2003, p. 509), multilevel DIF analysis enables the study of “a myriad of contextual variables at each level that are potentially related to DIF”. The amount of information gained via multilevel DIF frameworks is substantial.

There are statistical benefits as well. Much of the data collected in the social sciences has a nested structure. This is particularly true in a state assessment context, where students are nested within classrooms, schools, neighborhoods, and districts, etc. Multilevel DIF frameworks improve estimates of the relationship between latent traits and predictors, by considering both between and within cluster variation (Kamata & Cheong, 2007). Ignoring the nested structure of the data has been shown to result in substantial inflation of the Type I error rate under certain conditions in DIF analyses (French & Finch, 2010).

The remainder of this section will focus on three multilevel DIF frameworks: (1) the multilevel Mantel-Haenszel (MMH; French & Finch, 2013), (2) the multilevel SIBTEST (French & Finch, 2015), and (3) a multilevel Rasch model (Kamata, 2001). They are termed frameworks instead of methodologies as they typically represent a family of potential methodologies or models, which are appropriate under differing conditions.

There is additional terminology which is salient regards the nesting of the data. As an example, data in which items are nested within students which are nested within schools would consist of an item level model (items), a person level model (students) and a social-unit level model (schools). Generic terms are used throughout as researchers can determine what unit is appropriate for analysis (e.g. classrooms or districts instead of schools). The non-parametric approaches, the multilevel Mantel-Haenszel and SIBTEST, attend to data with only two-levels: person and social-unit. The parametric approach, the multilevel Rasch model, dictates two-levels: item and person, with the ability to extend to a third level, the social-unit level.

Multilevel Mantel-Haenszel. The multilevel Mantel-Haenszel (MMH; French & Finch, 2013, 2010) is a family of adjustments to the Mantel-Haenszel for multilevel data structures. Three possible adjustments have been proposed in the literature: (1) an adjustment to the ability estimates (Pommerich, 1995; Zhang & Boos, 1997), (2) an adjusted test statistic accounting for level-2 covariance (Begg, 1999), and (3) an adjustment based on a meta-analytic framework (Cooper & Hedges, 1994). While these adjustments to the Mantel-Haenszel are all twenty-plus years old, they are just beginning to garner attention in DIF literature.

Only one of the adjustments, that proposed by Begg (1999) will be presented due to its superior results in a simulation study undertaken by French and Finch (2013). As will be further detailed later, the Begg Mantel-Haenszel method (BMH) represents a series of possible adjustments. In their analysis, French and Finch (2013) utilize a purified scale score as the matching criterion, thus the matching criterion will be referred to as a subtest.

The BMH approach involves estimating the variance in the *MH CHISQ* statistic due to the clustering of examinees in addition to the “naïve” variance which assumes no such clustering. The *MH CHISQ* test statistic is then adjusted using a factor based on the ratio of the score statistic variance to the naïve variance of the score statistic. Logistic regression is used to estimate the variances for the naïve and clustered model:

$$\ln\left(\frac{P_{ki}}{1-P_{ki}}\right) = \beta_0 + \beta_1 X_i + \beta_2 Y_i, \tag{18}$$

where

P_{ki} is the probability of a correct response to item *k* by person *i*,
β_0 is the intercept,
X_i is the group membership for subject *i*,
Y_i is the matching subtest score for subject *i*,
β_1 is the coefficient for the grouping variable, and
β_2 is the coefficient for the matching subtest variable.

The model has an associated score statistic that tests the null hypothesis of no association between the predictor variables and the response. In the naïve model, the covariance matrix for the response with respect to clusters is the identity matrix, in which the off-diagonal elements are zero. Therefore, the intraclass correlation (ICC), or correlation of responses within a common cluster, is zero. The clustered model estimates the off-diagonal elements of the covariance matrix. The approach estimates unique covariance for each cluster. The variance of the score statistics are then obtained from the covariance matrices and a ratio determined:

$$f = \frac{\sigma^2_{clustered}}{\sigma^2_{naive}}, \tag{19}$$

The adjusted MH CHISQ statistic then takes the following form:

$$MH_B = \frac{MH\ CHISQ}{f}, \tag{20}$$

where *MH CHISQ* is the Mantel-Haenszel chi-square test statistic. When there is no correlation among examinees from the same social-unit (e.g., school), $MH_B = MH\ CHISQ$.

The BMH method suffers from relatively low power for DIF detection in a number of conditions which were determined by French and Finch (2013) prior to conducting their formal simulation study. Therefore, three variations are proposed, multiplying *f* by 0.85 (BMH85), 0.9 (BMH9), or 0.95(BMH95) to improve power while maintaining low Type I error rates (French & Finch, 2013). French and Finch make specific recommendations regarding which adjustment to use based on the magnitude of the ICC.

Multilevel SIBTEST. French and Finch (2015) also extended the single level SIBTEST (Shealy & Stout, 1993) and CSIBTEST (Li & Stout, 1996) for uniform and nonuniform DIF respectively to a multilevel framework. Comparable to the MMH, the multilevel SIBTEST and CSIBTEST represent a family of methods for considering the multilevel structure of the data. Three possible adjustments have been proposed: (1) the bootstrap standard error (BSSE) approach, (2) the bootstrap-*t* (BST) approach and (3) the Moulton correction (Moulton, 1986). For detecting uniform DIF in multilevel data the BSSE approach is recommended, while the BST approach is recommended for detecting nonuniform DIF in multilevel data. The Moulton correction suffered from low power rates under conditions of uniform and nonuniform DIF. Only the BSSE approach will be discussed as the focus of the current study is on uniform DIF.

Of note, no correction is recommended when the groups to be compared are based on a within cluster variable (person level), e.g. gender or ethnicity. In this case the traditional SIBTEST and CSIBTEST sufficiently controlled the Type I error rate and had higher power rates than the multilevel adjustments (French & Finch, 2015). These corrections are recommended when the groups to be compared are based on a between cluster variable (social-unit level), e.g. teacher tenure or classroom curriculum.

The process for the BSSE adjusted SIBTEST involves six steps (French & Finch, 2015):

1. Calculate the test statistic of interest ($\hat{\beta}_{UNI}$) for the original sample.
2. Resample *m* blocks of individuals with replacement, where *m* is equal to the number of social-unit blocks in the sample.
3. For each bootstrap sample calculate the parameter estimate using SIBTEST.
4. Repeat this procedure multiple times (*B*, e.g. 200).
5. Calculate the BSSE as

$$S_{sibtest} = \left(\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\beta}_b - \bar{\beta}\right)^2\right)^{1/2}, \qquad (21)$$

and

$$\bar{\beta} = \frac{\sum_{b=1}^{B}\hat{\beta}_b}{B}, \qquad (22)$$

where $\hat{\beta}_b$ is the SIBTEST statistic for bootstrap sample b, and B is the total number of bootstrap samples.

6. Use the BSSE ($S_{sibtest}$) to construct the test for DIF as

$$z = \frac{\beta_{UNI}}{S_{sibtest}}, \tag{23}$$

where β_{UNI} is the SIBTEST statistic for the original data. Z is compared against the standard normal distribution to determine statistical significance.

Two- and Three-Level Rasch Models.

Multilevel Rasch Models with Invariance. The Rasch model can be conceptualized as a two-level hierarchical generalized linear model (HGLM), where the level-1 model for the dependent variable Y_{ij}, which is the response of examinee *j* to item *i*, consists of: (1) a sampling model, (2) a link function, and (3) a structural model. The sampling model is the Bernoulli distribution where the expected value and variance for Y_{ij} are

$$E(Y_{ij}|p_{ij}) = p_{ij} \text{ and } \mathrm{Var}(Y_{ij}|p_{ij}) = p_{ij}(1 - p_{ij}), \tag{24}$$

where the probability of person *j* answering item *i* correctly is defined as p_{ij}. The link function is the logit link:

$$\log[p_{ij}/(1 - p_{ij})] = \eta_{ij}, \tag{25}$$

where η_{ij} is the log-odds of person *j* answering item *i* correctly.

The probability that person *j* answers item *i* correctly can be rewritten as

$$p_{ij} = \frac{\exp(\eta_{ij})}{1 + \exp(\eta_{ij})}. \tag{26}$$

The first-level describes how the item effects and person abilities shape the log-odds of a correct response. The structural model at level-1 is

$$\begin{aligned} \eta_{ij} &= \beta_{0j} + \beta_{1j}X_{1j} + \beta_{2j}X_{2j} + \cdots + \beta_{(k-1)j}X_{(k-1)j} \\ &= \beta_{0j} + \sum_{i}^{k-1} \beta_{ij}X_{ij}, \end{aligned} \tag{27}$$

where *i* represents items (*i*=1, ..., k) and *j* represents persons (*j* = 1, …, *n*). X_{ij} is a dummy variable for item *i* and person *j*. Traditionally, X_{ij} is coded as one when an item response represents the *i*th item and zero otherwise. Table 7 provides an example of how the data will need to be transformed and coded for proper analysis using a multilevel Rasch model.

Table 7

Example Dummy Coding for HGLM Rasch Model

Person	X_{1j}	$X_{2\,j}$	X_{3j}	X_{4j}	...	X_{ij}	Response
j	1	0	0	0	...	0	1
j	0	1	0	0	...	0	0
j	0	0	1	0	...	0	1
j	0	0	0	1	...	0	0
j	0	0	0	0	...	1	0

One item must be dropped from the design matrix of the model for the item-level model to be identifiable and this item is called the reference item. For this reason, items range from 1 to k-1. Selection of the reference item would ideally be an item of average difficulty. The coefficient β_{0j} is the intercept term and it represents the effect of the reference item. The coefficient β_{ij} is the difference of effect from β_{0j} and it is associated with X_{ij} when an item response is on the *i*th item, coded as 1.

The second level defines how the abilities vary over the population of examinees. The level-2 (person level) models are formulated as

$$\begin{gathered}\beta_{0j} = \gamma_{00} + \mu_{0j} \\ \beta_{1j} = \gamma_{10} \\ . \\ . \\ \beta_{(k-1)j} = \gamma_{(k-1)0}.\end{gathered} \tag{28}$$

In the first formula, γ_{00}, corresponds to the effect of the reference item and is a fixed component of β_{0j}; μ_{0j} is the random component of β_{0j}. It is distributed as $N(0, \tau)$ and represents the ability of person j. $\beta_{(k-1)j}$ is equal to $\gamma_{(k-1)0}$ and represents the deviation of the *i*th item's effect from the effect of the reference item. By formulating an unconditional model for the abilities and fixing the item effects, the IRT assumption of invariance is satisfied.

Assuming items are dummy coded as 1 or 0, when the level-1 and level-2 models are combined the linear predictor model becomes $\eta_{ij} = \gamma_{00} + \mu_{0j} + \gamma_{i0}$, for person *j* and item *i*. Then, the probability that persons *j* answers item *i* correctly can be rewritten as

$$p_{ij} = \frac{\exp(\eta_{ij})}{1+\exp(\eta_{ij})} = \frac{\exp\left[\mu_{0j} - (-\gamma_{i0} - \gamma_{00})\right]}{1+\exp\left[\mu_{0j} - (-\gamma_{i0} - \gamma_{00})\right]} = \frac{\exp(\theta_j - b_i)}{1+\exp(\theta_j - b_i)}. \qquad (29)$$

The above equation demonstrates the equivalency to the Rasch model, where $\theta_j = \mu_{0j}$ and $b_i = -\gamma_{i0} - \gamma_{00}$.

This comparison highlights a key difference between the Rasch model and multilevel Rasch model regarding parameter estimates. The estimates obtained from the multilevel Rasch model can be interpreted as item easiness estimates, where a larger value represents an easier item. In contrast, a larger value for the IRT difficulty parameter (b_i) represents a harder item. It is possible to use a coding scheme of -1 and 0 to result in difficulty estimates in keeping with IRT, although it is not often presented in the literature.

Kamata (2001) expanded his two-level hierarchical Rasch model to a three-level model. Level-1 remains the same as in the two-level iteration (eq 27) except an additional subscript, *m*, is added to indicate social-units. The level-1 (item level) model can be written as

$$\eta_{ijm} = \beta_{0jm} + \beta_{1jm}X_{1jm} + \beta_{2jm}X_{2jm} + \cdots + \beta_{(k-1)jm}X_{(k-1)jm} = \beta_{0jm} + \sum_{i}^{k-1} \beta_{ijm}X_{ijm}, \qquad (30)$$

where i=1, …, k-1, j=1, …, n, and m = 1, …,. r. X_{ijm} is the dummy variable for the ith item for person j within social-unit m. β_{0jm} is the effect of the reference item, and β_{ijm} is the effect of the ith item compared to the reference item.

The level-2 (person level) model for person j in social-unit m can be written as

$$\begin{gathered} \beta_{0jm} = \gamma_{00m} + \mu_{0jm} \\ \beta_{1jm} = \gamma_{10m} \\ . \\ \beta_{(k-1)jm} = \gamma_{(k-1)0m}. \end{gathered} \tag{31}$$

Again, this model is identical to the person level model under a two-level conceptualization (eq 28) except for the additional subscript m. Here, μ_{0jm} indicates how much person j in social-unit m deviates from the mean of μ_{0jm} for social-unit m, which is denoted as r_{00m}. The variance of μ_{0jm}within class is τ_Υ and is assumed to be identical for all social-units. Additionally, γ_{00m} is the effect of the reference item in social-unit m and γ_{i0m} is the deviation of the ith item's effect from the effect of the reference item.

The level-3 (social-unit level) model is written as

$$\begin{gathered} \gamma_{00m} = \pi_{000} + r_{00m} \\ \gamma_{10m} = \pi_{100} \\ . \\ . \\ \gamma_{(k-1)0m} = \pi_{(k-1)00}, \end{gathered} \tag{32}$$

where r_{00m} is distributed N(0,τ_π). At level-3, π_{000} is the fixed component of γ_{00m} and r_{00m} is the random component. The variance of r_{00m} is τ_π. The IRT assumption of invariance is upheld by restricting all other items, γ_{10m} to $\gamma_{(k-1)0m}$, to only fixed components.

The combined linear model is

$$p_{ijm} = \frac{\exp(\eta_{ijm})}{1+\exp(\eta_{ijm})} = \frac{\exp\left[(r_{00m}+\mu_{0jm})-(-\pi_{i00}-\pi_{000})\right]}{1+\exp\left[(r_{00m}+\mu_{0jm})-(-\pi_{i00}-\pi_{000})\right]} = \frac{\exp(\theta_j - b_i)}{1+\exp(\theta_j - b_i)}, \tag{33}$$

where $\theta_j = r_{00m} + \mu_{0jm}$ and can be considered the ability for person *j* in social-unit *m*. In a three-level model, ability is divided into two parts. First, r_{00m} is the random effect associated with social-unit *m* and can be interpreted as the average ability of students in social-unit *m*. Second, μ_{0jm} is the person-specific ability of person *j* in social-unit *m*, i.e. how much person *j* deviated from the average ability of students in social-unit *m*. Lastly, $b_i = (-\pi_{i00} - \pi_{000})$ and represents the item difficulty for the *i*th item, where π_{000} is the item difficulty of the reference item.

Multilevel Rasch Models as Frameworks for DIF. The two- and three-level models proposed by Kamata (2001) offer a tremendous amount of flexibility, including the ability to detect a variety of different DIF scenarios. Kamata and colleagues (2005) modified the three-level Rasch model by introducing a coefficient corresponding to person-level DIF and allowing that coefficient to be random across higher level clusters, in this case schools. The models used in their analyses are presented below.

The level-1 (item level) model is unchanged (eq 30). The level-2 (person level) model is expressed as

$$\begin{gathered}\beta_{0jm} = \gamma_{00m} + \gamma_{01m}G_{ijm} + \mu_{0jm} \\ \beta_{1jm} = \gamma_{10m} + \gamma_{11m}G_{1jm} \\ . \\ . \\ \beta_{(k-1)jm} = \gamma_{(k-1)0m} + \gamma_{(k-1)1m}G_{(k-1)jm},\end{gathered} \qquad (34)$$

where G_{ijm} is a group membership variable (focal or reference) at the student level. In this conceptualization, γ_{00m} represents the mean of the abilities. Therefore, γ_{01m} represents the main effect of group or the difference between the reference and focal group and is an indicator of impact. Finally, γ_{i0m} is the difficulty of the *i*th item for $G_{ijm} = 0$ and γ_{i1m} is the difference of item difficulty for $G_{ijm} = 1$. If γ_{i1m} is statistically significant or meaningfully large it is an indication of DIF for the *i*th item.

DIF parameters (γ_{i1m}) are treated as random effects and the level-3 (social-unit model) becomes

$$\begin{gathered}\gamma_{00m} = \pi_{000} + r_{00m} \\ \gamma_{01m} = \pi_{010} + r_{01m} \\ \gamma_{10m} = \pi_{100} \\ \gamma_{11m} = \pi_{110} + r_{11m} \\ . \\ . \\ \gamma_{(k-1)0m} = \pi_{(k-1)00} \\ \gamma_{(k-1)1m} = \pi_{(k-1)10} + r_{(k-1)1m},\end{gathered} \qquad (35)$$

where r_{i1m} is the random effect of DIF across social-units. If the variance of r_{i1m} is larger than 0, the DIF effect varies across schools. Therefore, the effect of the student-level group membership is different from school to school. Jak, Oort, and Dolan (2013) define this as cluster bias.

If the variance of r_{i1m} is significantly different from zero or meaningfully larger, further investigation is warranted. Kamata et al.'s (2005) study added a social-unit level predictor with the goal of reducing the variation of r_{i1m}. The incorporation of a social-unit level predictor, *GS*, results in the following level-3 model (school level),

$$\begin{gathered}
\gamma_{00m} = \pi_{000} + r_{00m} \\
\gamma_{01m} = \pi_{010} + r_{01m} \\
\gamma_{10m} = \pi_{100} \\
\gamma_{11m} = \pi_{110} + \pi_{111} GS_{11m} + r_{11m} \\
\cdot \\
\cdot \\
\gamma_{(k-1)0m} = \pi_{(k-1)00} \\
\gamma_{(k-1)1m} = \pi_{(k-1)10} + \pi_{(k-1)11} GS_{(k-1)1m} + r_{(k-1)1m},
\end{gathered} \tag{36}$$

where π_{i11} indicates the interaction of the student level grouping variable and the school level grouping variable. A significant *p*-value indicates that the student level effect is significantly different across social-units, dependent upon the school level grouping variable. The amount of reduction in the variance of r_{i1m} by including *GS* in the model can be evaluated by using their chi-square statistics and a likelihood-ratio test. Analyses of this nature align well with notions of contextualized item responses and DIF as presented in an ecological model.

Multilevel DIF Detection Literature

Studies using multilevel DIF frameworks have increased in recent years but the research remains relatively sparse. Much of the research on multilevel Rasch models has focused on applied studies (Burkes, 2009; Cai, 2015; Cheong, 2006; Cid, 2009; Kamata et al., 2005; Li et al., 2017) or limited the model to two-levels (Chen, Chen & Shih, 2013;

Lui, 2011; Zhu, Rupp & Gao, 2011). Comparative studies tend to compare multilevel DIF frameworks to their corresponding single level method (Acar, 2012; Atar, 2007). Currently, there is no existing research which compares multiple multilevel DIF approaches across families of frameworks. Literature concerning the non-parametric approaches, the multilevel Mantel-Haenszel and multilevel SIBTEST, will be presented first followed by literature regarding the parametric approach, the multilevel Rasch model. Lastly, the issue of anchor item selection and purification will be discussed. While not a central focus of this study, this is a key issue in DIF detection regardless of framework chosen.

Non-parametric Approaches: Multilevel Mantel-Haenszel and Multilevel SIBTEST. French and Finch (2013; 2015) conducted two identical simulation studies investigating the power and Type I error rates of the multilevel Mantel Haenszel and multilevel SIBTEST approaches. Though not comparative studies, both varied the sample size within clusters (person level sample size), the number of clusters (social-unit level sample size), the intraclass correlation (ICC), and the magnitude of DIF. Neither study investigated the effect of impact, the equivalency of the focal to reference group sample sizes, or the balance of the DIF items (heavily favoring one group versus evenly balanced). All factors which are known to affect the power and Type I error rates.

French and Finch (2013) recommend using three modifications to the BMH to increase power without sacrificing control over the Type I error rates. Specifically, multiplying the final statistic by 0.85 (BMH85), 0.90 (BMH9), and 0.95 (BMH95). When examining social-unit level DIF, the BMH85 method results in the highest power but also in inflated Type I error rates. If the intraclass correlation (ICC), or correlation of responses within a common cluster, is higher than 0.25 then the BMH85 is the preferable method due to the BMH85's higher power as all methods resulted in inflated Type I error. When the ICC is less than 0.25, either the BMH9 or BMH95 can be used based on the practitioner's tolerance for Type I error and desired power.

The observed power rates were largely dependent on the person level sample size, degree of within cluster correlation, and magnitude of DIF. Larger within cluster sample sizes are required to reach adequate power for ICC values of 0.25, 0.35 and 0.45 when the magnitude of DIF is 0.4 or 0.6. However, such high ICC values may not be reasonable in educational data (Hedges & Hedberg, 2007), lessening the sample size burden.

French and Finch (2015) conducted an identical simulation study to compare the multiple multilevel alternatives for the SIBTEST. When examining social-unit level DIF the BSSE approach yielded the highest power and best control of the Type I error rate. The multilevel SIBTEST BSSE approach was able to hold the Type I error rate below 0.05 across all conditions. Power was primarily influenced by the magnitude of DIF and the ICC, with smaller magnitude DIF and higher ICC values suffering from lower power.

Of note, neither study examined effect size and the multilevel DIF frameworks ability to accurately estimate the DIF effect size.

Parametric Approaches: Multilevel Rasch Model. While difficult to group studies using the HGLM framework because of model nuances, significant results from the literature will be presented. A discussion of estimation methods will be followed by general considerations of sample size and distribution. It is important to note that few multilevel Rasch DIF simulation studies focus on three levels (for examples, see: Binci, 2007; Kamata et al., 2005; Wen, 2014)

Multiple types of estimation methods are possible for the two- and three-level Rasch model. Penalized-quasi maximum likelihood (PQL) and Laplace approximation of maximum likelihood are the most widely used. Both methods are available in popular software (PQL in HLM, Laplace or PQL in SAS GLIMMIX). Raudenbush and colleagues (2000) demonstrated that the Laplace method when compared to PQL produced remarkably accurate parameter estimates. In a simulation study comparing the two estimation methods for a three-level Rasch model containing DIF, Binci (2007) found that the Laplace method generally outperformed the PQL method in larger cluster sizes regarding stability of estimates for the item difficulty parameter and DIF parameters, and the stability of variance estimates when regarding DIF as random effects. Type I error rates for the estimation methods were comparable as were their power rates for detecting DIF. Power rates were unsatisfactorily low, however, rates approached acceptable levels in the largest simulation conditions (40 person/cluster).

Sample size studies tend to converge on the conclusion that when examining social-unit level DIF, within cluster sample size should be kept to a minimum of 30 persons to maintain acceptable power rates (Binci, 2007; Wen, 2014; Zhu et al., 2011). However, these studies looked at relatively small social-unit sample sizes. While utilizing three-level models, Wen (2014) considered only 100 clusters, and Binci (2007) looked at 20, 30, 40, and 50 clusters. Zhu and colleagues (2011) considered only two-level models with a total sample size of 1,000 and short tests with a large percentage of DIF (12 items with four containing DIF). They observed power rates in the range of 0.50 to 0.70 under even moderate DIF conditions but their results are not directly comparable to Wen's and Binci's.

Impact has minimal effects on the power and Type I error rates within a two-level Rasch DIF model context (Zhu et al., 2011). However, impact has not been included in current simulation studies on three-level Rasch DIF models.

Anchor Item Selection and Purification. As presented above the Mantel-Haenszel and multilevel Mantel-Haenszel use all items in the test to serve as the matching variable and are therefore assuming items other than the studied item are DIF-free. This is referred to as the All-Other Anchor Method. However, as the number of DIF items within a test increases, the degree of violation of the assumption increases (Wang, Shih, & Sun, 2012). Numerous studies have demonstrated that scale purification procedures provide a substantial improvement over assuming all but the studied item are invariant (e.g. Clauser, Mazor, & Hambleton, 1993; Wang, 2004). Of note, in the only published research on the multilevel Mantel-Haenszel for DIF detection French and Finch (2013) make use of a purified subtest. When using a purified subtest, only items which are DIF free and the studied item are included.

The SIBTEST and multilevel SIBTEST make use of a subset of anchor items (valid subtest) which are assumed to be DIF free. This method uses the anchor items to establish a common metric for evaluating DIF in all the other items and is commonly referred to as the constant-item (CI) method (Wang & Yeh, 2003). Within the SIBTEST program, users specify anchor items. The CI method has been shown to yield high power with as few as four anchors in tests containing as many as 40% DIF items (Thissen et al., 1988; Wang & Yeh, 2003).

The multilevel context complicates anchor item selection, particularly in the parametric approach. Directly related to the issue of anchor selection is the quality of the reference item as it is not included in DIF analyses and is therefore part of the anchor set. Zhu and colleagues (2011) found that when the reference item contained DIF, Kamata's (2001) two-level Rasch model suffered from low power and extremely high Type I error rates. Therefore, it is imperative to select an appropriate item for use as the reference item.

Numerous approaches have been suggested for purification and can be classified into three categories: (1) model building, (2) purified subtest, and (3) anchor selection. Within the multilevel Rasch DIF framework, model building can be viewed as an umbrella which encompasses the purified subtest and anchor selection methods. Of note, the purified subtest and anchor selection methods detailed below were designed for two-level Rasch DIF models. While they are generalizable to three-levels they would require additional steps to ensure items deemed DIF-free were free of DIF at both the person and social-unit level if both levels were of interest.

Model-Building. As an approach for DIF detection, model building is only relevant to the multilevel Rasch model. In general, the model building process should progress through the following steps (Cheong, 2006):

1. Estimation of the unconditional model, i.e. the model without covariates at the person level and social-unit level.
2. Estimation of a conditional model with group membership as predictors at level-2.
3. Investigate and assess which items exhibit DIF and the patterns, directions, and magnitude of the detected DIF.
4. Include a social-unit level (level-3) variable in the DIF screening procedure and repeat Steps 2 and 3.

However, the picture presented by Cheong (2006) is overly simplistic. First, Cheong assumes the only scenarios of interest are those in which DIF is investigated at the person level with social-unit covariates included to explain variation in person level DIF. Second, within Steps 3 and 4, multiple decisions will need to be made regarding anchor item selection. The most simplistic view would be to select no anchor items and test all items but the reference item for DIF. This approach is referred to as exploratory (Wen, 2014). While simplistic, it suffers from low power. In a simulation study, Wen (2014) reported power levels ranging from 0.221 to 0.842, when the magnitude of level-3 DIF was 0.5 and 0.8, respectively. The other end of the spectrum involves confirmatory testing where only items known to contain DIF are tested. While this method has been shown to have higher power rates (Wen, 2014) it is an unreasonable assumption to assume researchers and practitioners will be able to pre-identify all DIF containing items accurately prior to analysis.

Chen, Chen, and Shih (2013), Cheong (2006), and Liu (2011) provide alternatives which fall within the spectrum of exploratory to confirmatory approaches. The most simplistic is Cheong’s suggestion that the exploratory approach be viewed as an omnibus test for DIF with a null hypothesis that all differences in difficulty parameter estimates between the focal and reference group are equal to zero. Should this hypothesis be rejected the researcher must determine which items to include in subsequent analyses. Cheong proposed including items where DIF (γ_{i1m}) was significant, p< 0.5, and the estimate of the group difference in item difficulty is equal to or larger than half a logit. These items are retained for further DIF analyses. All other items are retained but the grouping covariate is removed and they are treated as invariant.

However, the model building approach as described above does not address the need for initially selecting a reference item which is DIF free. The purified subtest and anchor selection approaches address this issue more fully. These approaches are also applicable to the multilevel Mantel-Haenszel and multilevel SIBTEST.

Purified Subtest. Purification can be defined as the process that removes the effect of the DIF items in the purified subtest so that DIF items can be accurately detected (French & Maller, 2007). For the Mantel-Haenszel, this results in a two-stage approach. First, all items are used to derive the observed score that is used as the matching criterion. In the second step, items with DIF are removed from the observed score. However, as noted by Zwick (1990) it is important to include the studied item in the matching criterion to avoid inflated Type I errors.

Liu (2011) addresses scale purification for DIF analysis using a two-level multilevel Rasch model. However, the proposed techniques would be applicable to the three-level Rasch model as well. Liu examined two approaches for deriving a purified subtest: the forward approach and the iterative approach. In the forward approach, all items are tested for DIF individually while all other items are considered part of the purified subtest. Then, all items which were flagged for significant DIF (α = 0.05) are tested simultaneously in a single model. All other items act as the purified subtest.

The iterative approach (Liu, 2011) proceeds similarly to the forward method. However, it is a two-stage procedure. In the first stage, all items are tested for DIF individually while all other items are considered part of the purified subtest. The subtest which is used in stage-two contains only those items which were not flagged for DIF in the first-stage. All items are tested again individually for DIF using the purified subtest. Lastly, all items which are flagged for DIF in stage-two are tested simultaneously in a single model with all other items acting as the purified subtest.

While the forward approach is more efficient, Liu (2011) concludes that in tests with high levels of contamination (40% DIF) the iterative approach is more powerful. The iterative procedure also had lower Type I error rates. Both procedures had power levels above 0.80 when *N*=2,000, across all other simulation conditions. The magnitude of DIF investigated was 0.60.

Anchor Selection. Utilizing anchor sets differs from purified subtests on one crucial aspect. Within anchor set methodologies, during the initial stage and all subsequent stages the same number of items will be used. This is not necessarily true in purified subtest methods. For example, assume items 1 and 2 of 10 have been flagged for DIF in stage-one of the iterative approach. In stage-two, items 1 and 2 would be tested for DIF using a matching variable of nine items (items 3-10, plus the studied item). However, in stage-two item 3 would be tested for DIF using a matching variable of only eight items (items 3-10) because items 1 and 2 are excluded from analyses unless they are the studied item.

Chen and colleagues (2013) recommend applying a “DIF-free-then-DIF” (DFTD; Wang, Shih, & Sun, 2012) purification approach to selecting an anchor item set when using the two-level Rasch model. While not directly applicable to the multilevel SIBTEST, SIBTEST and its multilevel extension also make use of anchor item sets. Within the DFTD, one method for selecting the anchor item(s) is to use a CI approach (H-IT method; Chen et al., 2013). In their study, the anchor set was specified to be either one or four items.

In the H-IT method, DIF should be assessed using the CI method *k* times by setting each item as the reference item iteratively (Chen et al., 2013). The absolute values of the DIF effects across iterations are averaged and the item(s) with the lowest absolute value are selected as the anchor set. All other items are then simultaneously tested for DIF.

An approach such as DFTD using the H-IT method is highly iterative, an alternative is selecting an anchor set arbitrarily in the first step and estimating DIF effects for all other items (H-OR method; Chen et al., 2013). In the second stage, the analysis is then rerun using the item(s) with the least value of DIF as the anchor set (either 1 or 4 items). This approach differs from the purified subtest procedures detailed above in that it does not make use of the All-Other anchor method but uses a predetermined number of anchor items in both stages one and two.

In regards to anchor set selection, Chen and colleagues (2013) defined power and Type I error in relation to the selection of a DIF free item as part of the anchor set. The H-IT method resulted in superior results when there was impact and when the anchor set consisted of four items. However, the H-OR method yielded superior results in terms of power and Type I error when there was no impact and the analysis was focused only on selecting a single anchor item. Chen and colleagues (2013) ultimately recommend use of the H-IT method in all cases due to its generally high-power rates and low Type I error rates across all conditions.

CHAPTER III

METHODS

This chapter has two sections. First, the design of the simulation study is presented including variables which were left as constant. Rationale for the decisions made is provided. Lastly, the evaluation criterion for these procedures is briefly described.

Simulation Design

Data were simulated for examinees on a dichotomously scored 40-item assessment. The multidimensional extension of the 2PL model (M2PL) was used for data generation. A multidimensional IRT model was chosen to replicate some of the noisiness experienced in real life data. The M2PL is given by

$$P\left(U_{ij} = 1 \middle| \theta_j, a_i, d_i\right) = \frac{e^{a_i\theta_j' + d_i}}{1 + e^{a_i\theta_j' + d_i}}. \qquad (37)$$

The *d* parameter is the intercept term and the elements of the *a*-vector are slope parameters (Reckase, 2009).

Data was generated from a M2PL with two dimensions. The first being the dimension of interest and the second being a nuisance dimension. The slope parameter of the ability dimension, a_1, was constrained to 1.0 for all invariant items. The second slope parameters, a_2, were constrained to zero for the thirty-four invariant items. The intercepts were drawn for a uniform distribution ranging from -2.0 to 2.0. For the six items containing DIF, a_1 was set to 0.8 and a_2 was set to 0.6, making the probability of a correct response dependent upon both dimensions. The intercepts were constrained to ensure the addition of DIF would not result in extreme item parameters. Item intercepts for the items which were manipulated to include DIF are included in Table 8.

Table 8

Item Intercepts for DIF Items

Item type	a_1	a_2	d_R	d_F
Low *b*				
1	0.8	0.6	1.50	1.50 + δ
2	0.8	0.6	1.50 + δ	1.50
Medium *b*				
3	0.8	0.6	0	0 + δ
4	0.8	0.6	0 + δ	0
High *b*				
5	0.8	0.6	-1.50	-1.50 + δ
6	0.8	0.6	-1.50 + δ	-1.50

Dimension one ability was generated from $N(\theta_{1j},1)$. where θ_{1j} is the social-unit cluster mean ability and is dependent upon simulation conditions. The second dimension ability, the nuisance ability, was generated from $N(0,1)$. The correlation between the first and second dimension, $\rho_{\theta_1\theta_2}$, was constrained to zero. This represents a somewhat overly simplistic scenario. The nuisance and ability dimension would likely be correlated. However, this represents a baseline scenario for testing the power and ability to estimate effect sizes for multilevel DIF frameworks.

A total of 64 conditions were analyzed using three DIF detection methods. One hundred replications within each condition were simulated. The conditions varied within the study are outlined in Table 9 and expanded upon below.

Table 9

Conditions Varied in the Study

Condition	Level
Social-Unit Sample Size	J= 100, 300
Intraclass Correlation	0.1, 0.2
Magnitude of DIF at the Social-Unit Level	δ = 0.0, 0.3, 0.5, 0.7
Impact	0.0, 0.5
Ratio of Reference to Focal Social-Unit Sample Size	1:1, 3:1

The conditions above are similar to those from several multilevel DIF studies. Research has explored several conditions that influence the impact of DIF detection when using multilevel frameworks. The conditions previously explored in studies which included nesting at the social-unit level include: person level sample size (Binci, 2007; French & Finch, 2013; French & Finch, 2015), social-unit sample size (Binci, 2007; French & Finch, 2013; French & Finch, 2015), intraclass correlation (French & Finch, 2013; French & Finch, 2015), the magnitude of DIF (French & Finch, 2013; French & Finch, 2015; Wen, 2014), the proportion of DIF items (Wen, 2014), and the equivalence of ratio and focal group sample size (Wen, 2014).

However, research on the effect of impact in multilevel DIF research is lacking. Binci (2007) examined impact but not in a comparative study, assuming an ability difference of -0.5 between the focal and reference group in all conditions. Additionally, only studies on the three-level Rasch model have considered the effect of the balance of DIF items and the equivalence of the ratio and focal group size. Within the three-level Rasch model DIF research, studies have neglected to systematically vary the ICC. Given the dearth of research on the effect of impact and focal to reference group ratio for the multilevel Mantel-Haenszel and SIBTEST and the effect of intraclass correlation for the three-level Rasch model, an emphasis was placed on including those factors within the current research.

DIF Detection Methods. Three methods for multilevel DIF detection where compared: (1) multilevel Begg Mantel-Haenszel method with an adjustment of 0.95 (BMH95), (2) multilevel SIBTEST using BSSE, and (3) three-level Rasch model. The first two methods were presented in Chapter Two. For the BMH, an adjustment 0.95 was chosen as ICC levels are kept below 0.25 per French and Finch's (2013) recommendations. The specific model used to test for social-unit level DIF within a multilevel Rasch framework is presented below.

Three-Level Rasch Model with Social-Unit Level DIF. The level-1 (item level) model is identical to that in an invariant three-level Rasch model (eq 31) and is written as

$$\begin{aligned}\eta_{ijm} &= \beta_{0jm} + \beta_{1jm}X_{1jm} + \beta_{2jm}X_{2jm} + \cdots + \beta_{(k-1)jm}X_{(k-1)jm} \\ &= \beta_{0jm} + \sum_{i}^{k-1} \beta_{ijm}X_{ijm},\end{aligned} \tag{38}$$

The level-2 (person level) model for person *j* in social-unit *m* is also identical (eq 32) and can be written as

$$\begin{gathered}\beta_{0jm} = \gamma_{00m} + \mu_{0jm} \\ \beta_{1jm} = \gamma_{10m} \\ . \\ . \\ \beta_{(k-1)jm} = \gamma_{(k-1)0m}. \end{gathered} \tag{39}$$

However, the level-3 (social-unit level) model is modified to include a social-unit level covariate which may be causing DIF. It is written as

$$\begin{gathered}\gamma_{00m} = \pi_{000} + r_{00m} \\ \gamma_{10m} = \pi_{100} + \pi_{101} GS_{10m} \\ . \\ . \\ \gamma_{(k-1)0m} = \pi_{(k-1)00} + \pi_{(k-1)01} GS_{(k-1)0m}, \end{gathered} \tag{40}$$

where $GS_{(k-1)0m}$ is the group membership at the social-unit level and $\pi_{(k-1)01}$ is the effect of the group membership at the social unit level, indicating DIF if significant. All other coefficients retain the same interpretation as in the invariant model presented in Chapter II.

Selected Software. This simulation made use of two statistical programming languages, SAS and R. All data was simulated in R using the mirt and mvtnorm packages and the author's own code. SIBTEST BSSE analyses were also conducted in R. Multilevel Rasch model and BMH95 analyses were conducted in SAS. Multilevel Rasch model analyses made use of proc glimmix while BMH95 analyses made use of code provided by French and Finch (2013). Aggregated results and plotting were conducted in R.

Variable Conditions.

Social-Unit Sample Size. Studies using three-level models have looked at the effect of varying the number of social-unit clusters (Binci, 2007; French & Finch, 2013; French & Finch, 2015). The results suggest that larger number of social-unit clusters result in higher power and lower Type I error for DIF detection. The multilevel Mantel Haensel and multilevel SIBTEST have demonstrated acceptable power levels with as few as 50 clusters under conditions of very large DIF (0.8) (French & Finch, 2013; French & Finch, 2015). However, the multilevel Rasch model had much lower power levels with 100 clusters and moderate DIF (0.5) even when a confirmatory approach was taken for DIF detection (Wen, 2014).

In the current study, the levels for social-unit sample size are *J*=100 or 200. These conditions were chosen as they seem to represent a realistic condition for an EOC or EOG assessment context. In a small state, these may represent utilizing all of the schools available while in larger state this may represent sampling from available schools. Additionally, investigating larger social-unit sample sizes will provide information regarding when the three-level Rasch model becomes adequately powered for detecting social-unit DIF.

Intraclass Correlation. Research from large national databases suggests the ICC values greater than 0.25 are rare in educational data (Hedges & Hedberg, 2007). Despite that finding researchers have investigated ranges from 0.05 to 0.45 (French & Finch, 2013; French & Finch, 2015). For non-parametric methods, as ICC values increase power generally decreases though to a lesser degree for moderate to large magnitude DIF

(0.6 and 0.8). In the current study, ICC values of 0.10 and 0.20 were chosen to mimic realistic levels in educational data.

Magnitude of Social-Unit DIF. Simulation studies investigating social-unit DIF detection have used DIF magnitudes ranging from 0.4 to 0.8. However, multilevel studies on real data have found smaller shifts in the difficulty parameter. Few real data studies have looked exclusively at social-unit level DIF, therefore results from person level studies are examined as well. For items flagged as having significant DIF, Kamata et al. (2005) found absolute differences in the difficulty parameter ranging from 0.240 to 0.541 at level-2. Cheong (2006) did not report difference in item parameter estimates, but of 13 items only flagged three as having more than a difference of 0.5. Liu (2011) found absolute differences ranging from 0.230 to 0.900 in a sample with a large difference in ability estimates, 1.22. Using TIMSS data, Burkes (2009) found absolute differences ranging from 0.45 to 0.60. Also using TIMSS data, Cai (2015) found absolute differences ranging from 0.09 to 0.52 when using a two-level Rasch model.

As the magnitude of DIF found in real data analysis was in general smaller than that observed in simulation studies, lower values were chosen for the current study. Specifically, 0.0 (null), 0.3 (small), 0.5 (medium), and 0.7 (large).

Impact. Mean ability differences, or impact, are common between groups in real data (Jodoin & Gierl, 2001). Impact has been found in real data studies using multilevel DIF detection methods (Liu, 2012). However, the current literature on social-unit DIF lacks examination into the effect of impact. At the person level, Zhu et al. (2011) found the two-level Rasch model to perform similarly in terms of power and Type I error when impact was present. The current study will investigate two conditions: no impact, and a mean ability difference of -0.5 between the focal and reference group. A value of -0.5 is similar to values observed in actual applications (Donoghue, Holland, & Thayer, 1993) and that used in the simulation study conducted by Zhu et al. Impact was introduced into the first dimension, the ability dimension, only.

Ratio of Reference to Focal Group Social Unit Sample Size. Research on the equivalency of the reference and focal group sample size when considering social-unit DIF is limited to studies utilizing the multilevel Rasch model. Wen (2014) demonstrated that when DIF is present only at the social-unit level higher power and lower Type I error rates were observed for a three-level Rasch DIF model when the reference to focal group ratio was equal. The current study investigated two conditions: equivalent sample sizes, and a reference to focal group ratio of 3:1.

Conditions Held Constant.

Within Cluster Sample Size. The within cluster sample size was held to *N*=20. While larger sample sizes have been demonstrated to result in higher power (Binci, 2007; French & Finch, 2013; French & Finch, 2015), a smaller sample size was deemed more realistic in an educational context. For instance, House Bill 13 in North Carolina caps third grade classrooms at 17 students with a three-student maximum buffer allowed. While higher grades are not subject to the same requirements, the highest average class size in North Carolina was 29 in grades 10-12 (Levinson & Schauss, 2017). According to a Schools and Staffing Survey (SASS) conducted by the National Center for Education

Statistics (2012) the average class size across grades ranges from 16.7 to 26.2 students in the United States.

Test Length. Multilevel DIF studies have ranged from very short tests (10 items; Zhu et al., 2011) to longer tests (40 items; French & Finch, 2013; French & Finch, 2015; Wen, 2014). According to Liu (2011), it is common practice to have 20 to 50 items in large-scale assessments. The most recent publicly available data placed North Carolina EOCs and EOGs anywhere from 52 to 75 items including field test items (NCDPI, 2013). For this study a test of 40 items was selected.

Proportion of Items with DIF. As the number of DIF items within a test increases, the power and Type I error rates can be negatively impacted (Wang, Shih, & Sun, 2012). However, purification and anchor selection techniques can mitigate this effect. For the current study, approximately 12% of the test was simulated to contain DIF (6 of 40 items). Additionally, anchor item selection techniques have been shown to be successful with DIF as high as 40% (Thissen et al., 1988; Wang & Yeh, 2003). Thus, it is realistic to make assumptions regarding the quality of the anchor set to be used under this degree of contamination.

Balance of Items with DIF. DIF items were manipulated so that the DIF achieved was balanced. Therefore, three items favored the reference group and three items favored the focal group. A dominant approach where DIF favors one group consistently is typically used in the multilevel context (e.g. French & Finch, 2013; French & Finch, 2015; Wen, 2014). Within the multilevel Rasch model literature, purified subtest and anchor selection procedures successfully selected a purified matching criterion (accuracy greater than 0.90) under balanced DIF conditions (Chen et al., 2013; Liu, 2012)

Selection Procedures for the Purified Subtest and Anchor Items. How the matching criteria for examinees was selected is dependent upon the method used for DIF detection. When the multilevel Mantel-Haenszel, BMH95, was used a purified subtest was assumed. In order to ensure the resulting analysis were not conflated by a contaminated subtest, the matching criterion consisted of the 34 non-DIF items when the non-DIF items were the studied item. When one of the DIF items was the studied item the matching criterion consisted of the 34 non-DIF items plus that item.

When DIF analyses were conducted for the multilevel SIBTEST BSSE and three-level Rasch model, an anchor item approach was utilized. The length of the anchor set was dependent upon the method used. For the SIBTEST BSSE, an anchor item set of twenty DIF free items was constructed. The remaining fourteen DIF-free items and six DIF-contaminated items served as the suspect group and were tested for DIF. Using DIMTEST to select unidimensional anchor sets across various conditions, Scott (2014) found that roughly 60% of items were selected. Thus, this percentage was selected for use with SIBTEST BSSE as it represents a more realistic and less confirmatory approach than that taken in previous multilevel literature (French & Finch, 2015).

When the multilevel Rasch model was used to detect DIF, the anchor item set contained four items. Studies have shown four-item anchor sets can be selected rather accurately across various conditions in a multilevel Rasch model context (Chen et al., 2013). Additionally, four-item anchor sets have been proposed and studied in single level studies (Shih & Wang; Woods, 2009). As with the multilevel Mantel-Haenszel analyses, a four-item anchor set of DIF free items was chosen to ensure resulting analyses were not conflated by a contaminated anchor set. The four items were selected at random from chosen ranges in order to ensure they represented the full range of the item intercepts, with an average of zero. All remaining items were then tested for DIF simultaneously.

Addressing the Research Questions

Given the simulation conditions described above, the dependent variables are the Type I error rate, the statistical power of the various multilevel DIF detection frameworks, and the resulting effect size measures. Type I error rate is the proportion of replications that the multilevel DIF detection method flags an item for DIF when an item does not contain DIF. Statistical power is the proportion of replications that the multilevel DIF detection method flags an item for DIF when the item has been defined to contain DIF. Effect size is the magnitude of DIF and can be conveyed in numerous ways, including as a log-odds ratio, on the ETS delta scale, or as the difference between the reference and focal group difficulty parameters.

The first research question investigates the empirical Type I error and power rates for the three main DIF frameworks. Type I error rates should be approximately 0.05 when the significance level of α=0.05 is used. Acceptable power rates were set at 0.80 (Cohen, 1988). The multilevel DIF framework which yields the highest statistical power and controls the Type I error rate, after considering the margin of error, will be the preferred method.

Two ANOVAs were conducted where the Type I error and power rates averaged across replications for each combination of conditions served as the dependent variables and the manipulated factors were the independent variables. This analysis addresses the second research question regarding which simulation conditions have the greatest impact on power and Type I error rates and is recommended in simulation research (Paxton, Curran, Bollen, Kirby & Chin, 2001).

The third research question pertains to the accuracy of effect size estimates under each multilevel DIF detection framework. Effect size estimates as logit differences between the reference and focal group were compared to the true values. Bias and the root mean square error (RMSE) between estimated and true effect sizes were calculated. Bias is the deviation between the estimated effect size and the true effect.

$$Bias = \frac{\sum_{i=1}^{N}(\hat{\delta}_i - \delta_i)}{N}, \tag{40}$$

where $\hat{\delta}_i$ is the estimated effect size, δ_i is the true effect size, and *N* is the replications in the simulation study. Bias estimates the distance from the estimated to true effect size as well as direction. The root mean square error (RMSE) is defined as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{\delta}_i - \delta_i)}{N}}. \tag{41}$$

RMSE evaluates the absolute magnitude of difference between the estimated and true effect size.

In order to calculate the bias and RMSE for the effect size and compare across multilevel methods, the measures of effect size must be placed on the same scale. The following relationships exist amongst the different measures of effect size under the Rasch model:

$$MH\ D - DIF = -15 * \hat{\beta}_{UNI} = -4(b_F - b_R). \tag{42}$$

The statistics for the multilevel Mantel-Haenszel methods and multilevel SIBTEST BSSE will be converted to differences in the difficulty between the focal and reference compare in order to be compared to the three-level Rasch model estimates and initially simulated differences.

CHAPTER IV

RESULTS

This chapter is divided into two sections, first the results of the simulation study will be presented. Then the results will be discussed in relationship to multilevel validation and the two frameworks which were presented (the multilevel IUA and ecological model of validation).

Descriptive Statistics

The descriptive statistics for each simulation condition were checked to ensure that the characteristics of the resulting data match those laid out in the previous chapter. Across conditions, the simulated data matched the targets set out in the simulation plan. Tables 10 and 11 include the relevant descriptive statistics.

Table 10

Descriptive Statistics for Intraclass Correlation

	Low ICC		High ICC	
	Mean	Standard Deviation	Mean	Standard Deviation
Null (δ=0.0)	0.10	0.01	0.20	0.01
Small (δ=0.2)	0.10	0.01	0.20	0.01
Medium (δ=0.4)	0.10	0.01	0.21	0.01
Large (δ=0.6)	0.10	0.01	0.21	0.01

Table 11

Descriptive Statistics for School Means

		Reference Ability Dimension Theta		Focal Ability Dimension Theta	
		Mean	Standard Deviation	Mean	Standard Deviation
Null (δ=0.0)	Equivalent	0.00	0.11	-0.01	0.15
	Impact	0.00	0.11	-0.51	0.15
Small (δ=0.2)	Equivalent	0.00	0.10	0.00	0.15
	Impact	0.00	0.10	-0.49	0.13
Medium (δ=0.4)	Equivalent	0.00	0.11	0.01	0.14
	Impact	0.00	0.11	-0.51	0.13
Large (δ=0.6)	Equivalent	0.00	0.11	-0.01	0.14
	Impact	0.00	0.11	-0.49	0.16

Power and Type I Error

Type I error and power rates were evaluated across conditions. To determine which manipulated factors influenced the power and Type I error rates, two ANOVAs were conducted. The average Type I error and power rates across replications for each combination of conditions served as the dependent variables and the manipulated factors were the independent variables.

Type I Error. The Type I error rate is the proportion of replications that the multilevel DIF detection method flags an item for DIF when an item does not contain DIF. Type I error rates near the level of chance are expected. When, $\alpha = 0.05$, acceptable Type I error rates fall between $\alpha \pm 1/2\ \alpha$. Therefore, Type I error rates between 0.025 and 0.075 will be considered reasonable (Bradley, 1978). Aggregate level results are presented in Table 12-15. Average Type I error rates broken out by condition are presented in Tables 16-18. The three multilevel DIF detection frameworks are presented separately due to space constraints. Rows represent the magnitude of DIF and columns represent all other manipulated conditions. Across all conditions and multilevel DIF frameworks, 77% of the Type I error results fell within the range considered reasonable.

Table 12

Average Type I Error by Magnitude of DIF

Magnitude of DIF	Average Type I Error
Null (δ=0.0)	0.043
Small (δ=0.2)	0.043
Medium (δ=0.4)	0.044
Large (δ=0.6)	0.044

Table 13

Average Type I Error by Social-unit Sample Size

Number of Clusters	Average Type I Error
J=100	0.048
J=300	0.040

Table 14

Average Type I Error by ICC

ICC	Average Type I Error
0.10	0.042
0.20	0.045

Table 15

Average Type I Error by Multilevel DIF Framework

Multilevel DIF Framework	Average Type I Error
BMH95	0.044
Multilevel Rasch Model	0.064
SIBTEST BSSE	0.022

When the BMH95 was used, acceptable Type I error rates were observed across conditions. When multilevel Rasch model was used, acceptable Type I error rates were observed under the low ICC condition (ICC=0.10). Under the high ICC condition (ICC=0.20), acceptable results were observed when the ability distributions for the focal and reference group were equivalent. When impact was present, the average Type I error rate was inflated. Use of the multilevel SIBTEST resulted in low Type I error rates across all conditions.

Table 16

BMH95 Type I Error

		High (ICC = 0.2)				Low (ICC = 0.1)			
		None (Δ=0.0)		Impact (Δ=0.5)		None (Δ=0.0)		Impact (Δ=0.5)	
		Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)
Null									
(δ=0.0)	100	0.045	0.039	0.044	0.043	0.038	0.043	0.042	0.042
	300	0.045	0.046	0.040	0.048	0.045	0.043	0.045	0.039
Small									
(δ=0.2)	100	0.046	0.046	0.039	0.048	0.045	0.041	0.041	0.044
	300	0.049	0.047	0.042	0.040	0.046	0.047	0.043	0.046
Medium									
(δ=0.4)	100	0.044	0.042	0.043	0.034	0.044	0.041	0.045	0.039
	300	0.044	0.050	0.049	0.042	0.041	0.044	0.048	0.045
Large									
(δ=0.6)	100	0.043	0.046	0.050	0.041	0.047	0.042	0.043	0.046
	300	0.046	0.039	0.051	0.049	0.044	0.040	0.047	0.053

Table 17

Multilevel Rasch Model Type I Error

		High (ICC = 0.2)				Low (ICC = 0.1)			
		None (Δ=0.0)		Impact (Δ=0.5)		None (Δ=0.0)		Impact (Δ=0.5)	
		Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)
Null									
(δ=0.0)	100	0.052	0.040	0.076	0.064	0.042	0.045	0.057	0.053
	300	0.045	0.051	0.130	0.137	0.051	0.047	0.056	0.054
Small									
(δ=0.2)	100	0.052	0.044	0.068	0.083	0.057	0.049	0.049	0.053
	300	0.053	0.045	0.117	0.120	0.049	0.046	0.055	0.067
Medium									
(δ=0.4)	100	0.046	0.049	0.082	0.057	0.052	0.049	0.046	0.048
	300	0.047	0.052	0.157	0.118	0.041	0.053	0.063	0.068
Large									
(δ=0.6)	100	0.046	0.041	0.087	0.071	0.047	0.046	0.054	0.058
	300	0.050	0.048	0.156	0.135	0.046	0.042	0.071	0.065

Table 18

SIBTEST BSSE Type I Error

		High (ICC = 0.2)				Low (ICC = 0.1)			
		None (Δ=0.0)		Impact (Δ=0.5)		None (Δ=0.0)		Impact (Δ=0.5)	
		Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)
Null									
(δ=0.0)	100	0.043	0.044	0.046	0.043	0.051	0.039	0.041	0.044
	300	0.001	0.000	0.000	0.001	0.000	0.001	0.000	0.001
Small									
(δ=0.2)	100	0.051	0.049	0.045	0.046	0.039	0.036	0.046	0.044
	300	0.001	0.001	0.001	0.000	0.001	0.001	0.001	0.000
Medium									
(δ=0.4)	100	0.046	0.037	0.046	0.055	0.054	0.041	0.043	0.041
	300	0.001	0.001	0.001	0.000	0.001	0.001	0.001	0.000
Large									
(δ=0.6)	100	0.051	0.032	0.039	0.047	0.043	0.035	0.053	0.041
	300	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000

ANOVA Results. A full factorial ANOVA was used to determine which of the manipulated variables and interactions significantly affected the Type I error rate. Full results are presented in Appendix A, Table A.1**.** ANOVA results which were statistically significant and resulted in at least a medium effect size (η^2>0.05) will be discussed. The highest order significant interaction which met the effect size requirement were two-way interactions. Because higher order interactions were significant main effects will not be presented. The two significant interactions which met the effect size criteria will be presented in the order of effect size magnitude.

The first was the social-unit sample size (*J*=100 or 300) by the multilevel DIF framework, $F_{(2,6)}$=912.03, *p*<0.001, η^2=0.117. Figure 8 includes the Type I error rate for a between cluster variable by social-unit sample size and multilevel DIF framework. Of note, under the condition of *J*=300, the SIBTEST BSSE method had negligible Type I error, below what would be expected by chance. These results are comparable to the Type I error rates reported by French and Finch (2015). The SIBTEST BSSE was noted to have decreasing Type I error rates when increasing the number of clusters from 100 to 200. Though only presently graphically the Type I error rates when the within cluster sample sizes are most similar to the condition used in this sample, *N*=15 and *N*=25, the Type I error rate appears to be near zero.

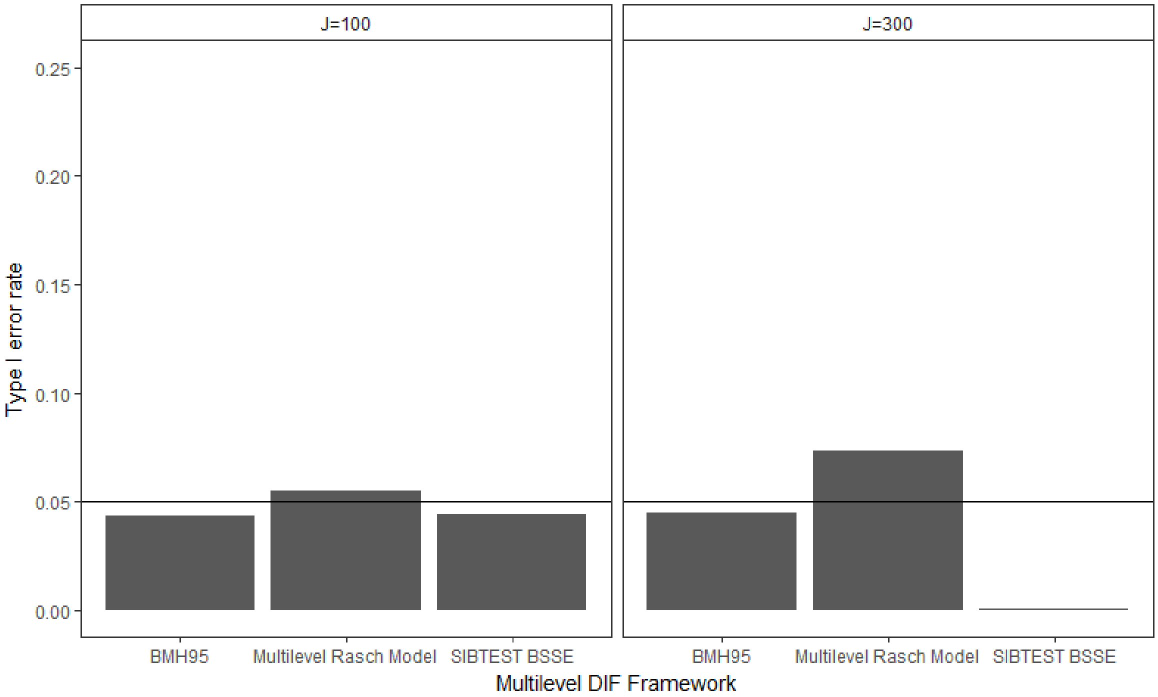

Figure 8. Type I Error for DIF Detection by Multilevel DIF Framework and Social-unit Sample Size.

Under both conditions, *J*=100 and *J*=300, the multilevel Rasch model resulted in Type I error above what would be expected by chance. However, when averaged across all other conditions the Type I error rate is below 0.075 and thus considered reasonable. The BMH95 and SIBTEST BSSE resulted in acceptable Type I error rates across both conditions.

The second was ability distribution by multilevel DIF framework, $F_{(2,6)}$=274.18, p<0.001, η^2=0.062. Figure 9 includes the Type I error rate for a between cluster variable by ability distribution and multilevel DIF framework. These results demonstrate that all methods maintained the Type I error rate when the ability distribution between the focal and reference groups were even. When impact was present, the multilevel Rasch model yielded inflated Type I error rates. However, the average Type I error rate is close to what would be considered reasonable. The Type I error rates for the BMH95 and SIBTEST BSSE methods are largely unaffected by impact.

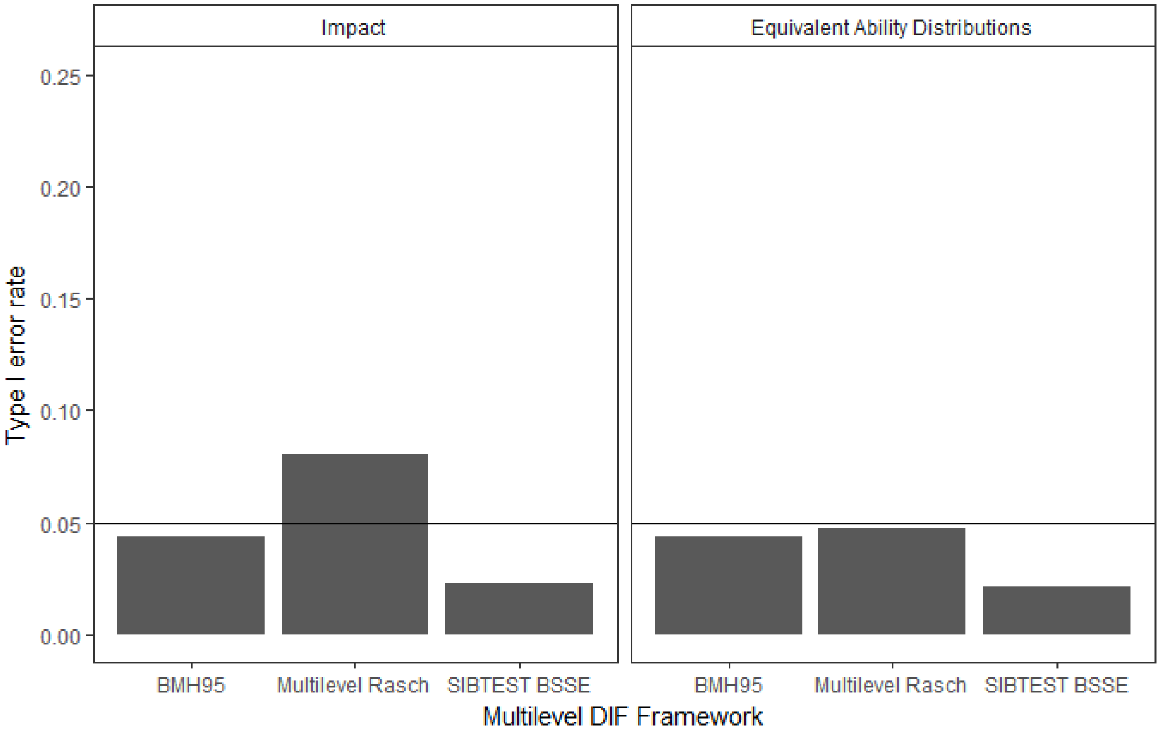

Figure 9. Type I Error for DIF Detection by Multilevel DIF Framework and Impact Presence.

Power. Statistical power is the proportion of replications that the multilevel DIF detection method flags an item for DIF when the item has been manipulated to contain DIF. Ideally power would be higher than 0.80, with the margin of error power rates of 0.782 will be considered highly powered. Aggregate results across conditions are present in Tables 19-22. Average power rates across all conditions and items are presented in Tables 23-25**.** As with the Type I error rates, rows represent the magnitude of DIF and columns represent all other manipulated conditions. The three multilevel DIF detection frameworks are presented separately due to space constraints.

Table 19

Average Power by Magnitude of DIF

Magnitude of DIF	Average Power
Small (δ=0.2)	0.740
Medium (δ=0.4)	0.974
Large (δ=0.6)	0.999

Table 20

Average Power by Social-Unit Sample Size

Number of Clusters	Average Power
J=100	0.854
J=300	0.956

Table 21

Average Power by ICC

ICC	Average Power
0.10	0.908
0.20	0.901

Table 22

Average Power by Multilevel DIF Framework

Multilevel DIF Framework	Average Power
BMH95	0.925
Multilevel Rasch Model	0.918
SIBTEST BSSE	0.871

When the BMH95 and multilevel Rasch model were used, high power rates were observed under all but the small magnitude DIF condition with a sample size of 100 clusters. However, the power rates for the multilevel Rasch model must be interpreted with caution as inflated error was observed when impact was present in the high ICC condition. When the multilevel SIBTEST was used, high power was observed for medium to large magnitude DIF. Under the condition of small DIF, acceptable power rates were only observed in the low ICC condition when there was an even ratio of reference to focal clusters and a sample size of 300 clusters.

Table 23

BMH95 Power

		High (ICC = 0.2)				Low (ICC = 0.1)			
		None (Δ=0.0)		Impact (Δ=0.5)		None (Δ=0.0)		Impact (Δ=0.5)	
		Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)
Small									
(δ=0.2)	100	0.682	0.557	0.677	0.532	0.717	0.542	0.678	0.560
	300	0.993	0.955	0.987	0.958	0.992	0.960	0.988	0.958
Medium									
(δ=0.4)	100	0.980	0.938	0.978	0.923	0.983	0.953	0.972	0.943
	300	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Large									
(δ=0.6)	100	0.983	0.983	1.000	1.000	1.000	1.000	1.000	1.000
	300	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 24

Multilevel Rasch Model Power

		High (ICC = 0.2)				Low (ICC = 0.1)			
		None (Δ=0.0)		Impact (Δ=0.5)		None (Δ=0.0)		Impact (Δ=0.5)	
		Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)
Small									
(δ=0.2)	100	0.670	0.560	0.657	0.530	0.697	0.540	0.667	0.558
	300	0.983	0.930	0.948	0.895	0.987	0.960	0.973	0.933
Medium									
(δ=0.4)	100	0.955	0.930	0.967	0.900	0.983	0.945	0.962	0.933
	300	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Large									
(δ=0.6)	100	0.998	0.995	1.000	0.997	1.000	1.000	1.000	0.998
	300	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 25

SIBTEST BSSE Power

		High (ICC = 0.2)				Low (ICC = 0.1)			
		None (Δ=0.0)		Impact (Δ=0.5)		None (Δ=0.0)		Impact (Δ=0.5)	
		Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)
Small									
(δ=0.2)	100	0.695	0.527	0.652	0.480	0.733	0.560	0.688	0.543
	300	0.798	0.528	0.773	0.523	0.828	0.620	0.773	0.548
Medium									
(δ=0.4)	100	0.970	0.888	0.982	0.908	0.990	0.933	0.982	0.923
	300	0.998	0.983	1.000	0.988	1.000	0.997	1.000	0.985
Large									
(δ=0.6)	100	1.000	0.998	0.998	0.998	1.000	0.995	1.000	0.992
	300	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

ANOVA Results. A full factorial ANOVA was also used to determine which of the manipulated variables and interactions significantly affected the power rate. Full results are presented in Appendix A, Table A.2. Similar to the ANOVA model predicting Type I error results, all manipulable conditions were included as factors predicting the power rates (model 1). However, a second full factorial ANOVA was run which also included which group the item favored in addition to the simulation conditions as an independent variable in the model (model 2). The full results are included in Appendix A, Table A.3.

Due to the nature of some of the simulation conditions, such as impact which favored the reference group, it was hypothesized that the power rates may differ depending on which group the DIF favored. A comparison of the two ANOVA models indicated that the more complex model did lead to a significantly improved fit over model 1 (p=0.003). Results for the more complex model will be presented. However, it should be noted that the group favored by the item never resulted in interactions or a main effect that were significant and met the criteria for a small effect size.

None of the higher-order interactions were statistically significant and resulted in a medium or larger effect size. As a result, interactions which had a small effect size will be presented. The highest order significant interaction which met the effect size requirement was the three-way interaction between DIF magnitude, social-unit sample size, and the multilevel framework. Because higher order interactions were significant main effects will not be presented. The results will be presented in two parts. First, the two-way interaction between magnitude of DIF and multilevel framework in order to highlight the general trend across models. Then, the three-way interaction will be presented.

The interaction between magnitude of DIF and social-unit sample size was significant, $F_{(2,4)}$=4,558.90, p<0.001, η^2=0.019. Figure 10 depicts the power rate for the magnitude of DIF by social-unit sample size. When the number of clusters is 300, the average power across all other conditions is acceptably high, even in the smallest DIF condition. However, when there are only 100 clusters the power level dips below acceptable levels for the small DIF condition. Indicating that at smaller social-unit sample sizes, on average the multilevel DIF frameworks were unable to adequately detect small magnitude DIF.

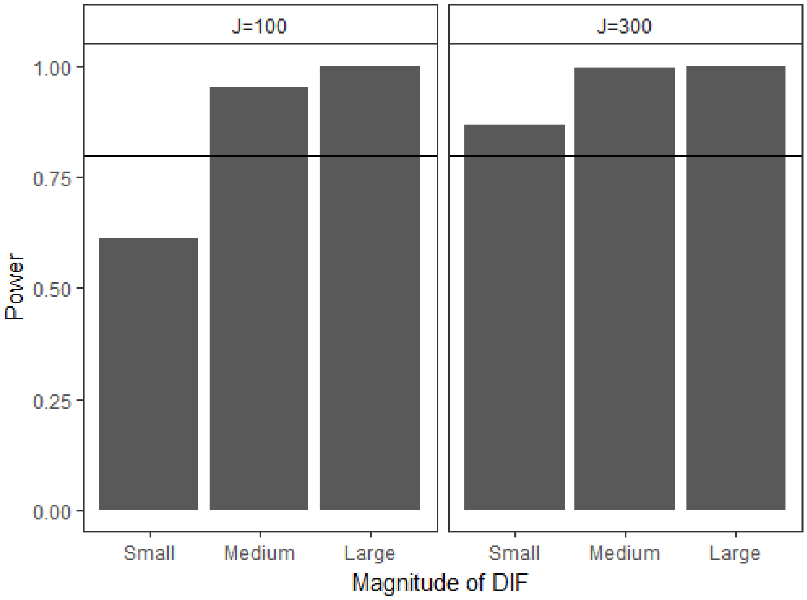

Figure 10. Power for Magnitude of DIF by Number of Social-unit Clusters.

The interaction between magnitude of DIF, social-unit sample size and the multilevel DIF framework was the highest-order significant interaction, $F_{(4,4)}$=7,142.77, *p*<0.001, η^2=0.013. Figure 11 depicts the power rate for the magnitude of DIF by social-unit sample size by multilevel DIF framework. As depicted in Figure 11, under the conditions of medium and large magnitude DIF, all multilevel DIF frameworks maintain adequate power. When the magnitude of DIF is small and the cluster size is 100, on average none of the multilevel DIF frameworks achieve an adequate level of power. However, under the small magnitude DIF condition when the cluster size is 300, the BMH95 and multilevel Rasch model maintain adequate power while the SIBTEST BSSE does not. Further investigation revealed that the power levels for SIBTEST BSSE reached an acceptable level when the focal to reference group was even under the condition of small magnitude DIF.

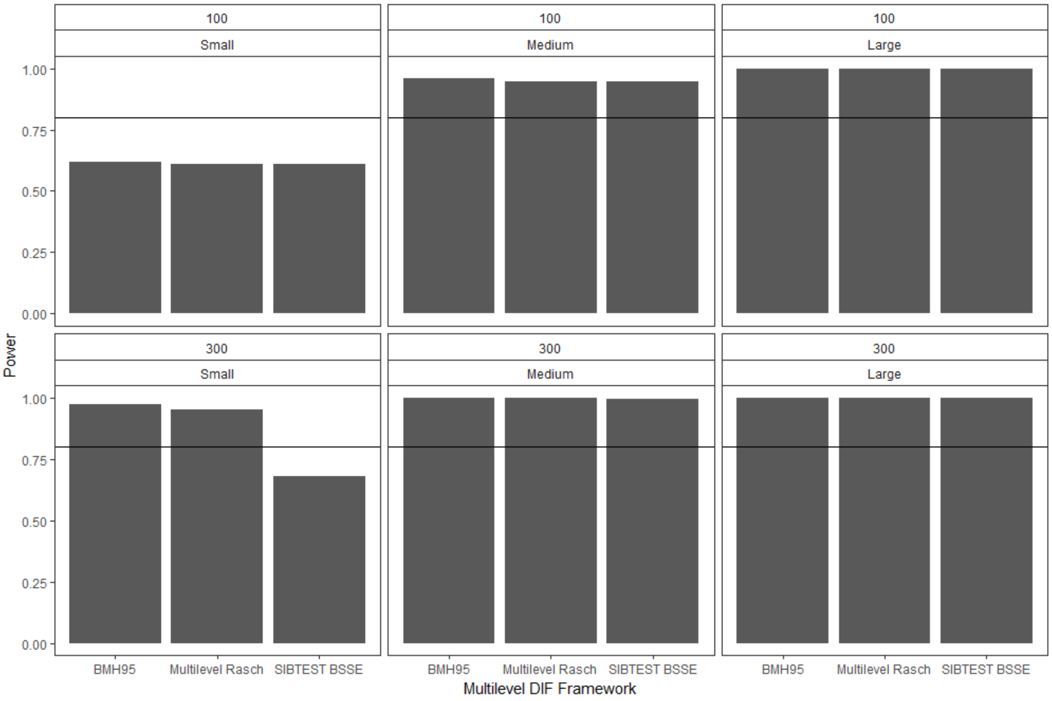

Figure 11. Power by Multilevel DIF Framework, Social-unit Sample Size, and Magnitude of DIF.

Effect Size Estimates

The BMH95 and SIBTEST BSSE only adjust the variance of the DIF detection method and therefore the statistical test for significance. As a result, research has focused on the Type I error and power of these methods. However, an understanding of how effect size estimates are impacted by multilevel data is a necessary component to understanding the operational usability of these frameworks. The effect size estimates were evaluated in terms of their relative bias and root mean square error.

Relative Bias. The relative bias estimates for the effect size are presented in Figures 12 through 16. Two vertical lines are plotted at -0.05 and 0.05 to outline the acceptable range of values. The bias values across conditions are included in Appendix B, Tables B1 through B6. Across all conditions the average relative bias was 0.001 with a standard deviation of 0.02. Indicating that on average the effect size estimates were relatively unbiased.

Both SIBTEST and SIBTEST BSSE effect size measures were retained during the study. However, as the bootstrapping procedure only adjusts the statistical test for significance and does not dramatically affect the SIBTEST parameter itself it was found that the estimated results were similar. For continuity with prior analyses, the effect size estimates using the averaged SIBTEST parameter from the bootstrapping procedure, i.e. the SIBTEST BSSE effect size estimate, will be presented.

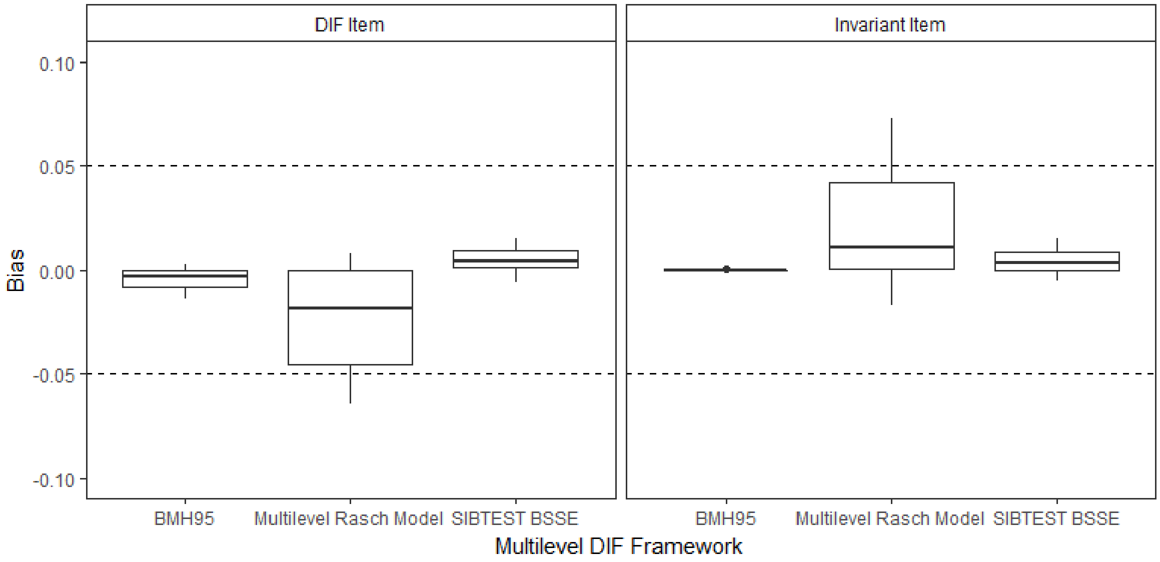

Figure 12. Relative Bias by Item Type and Framework.

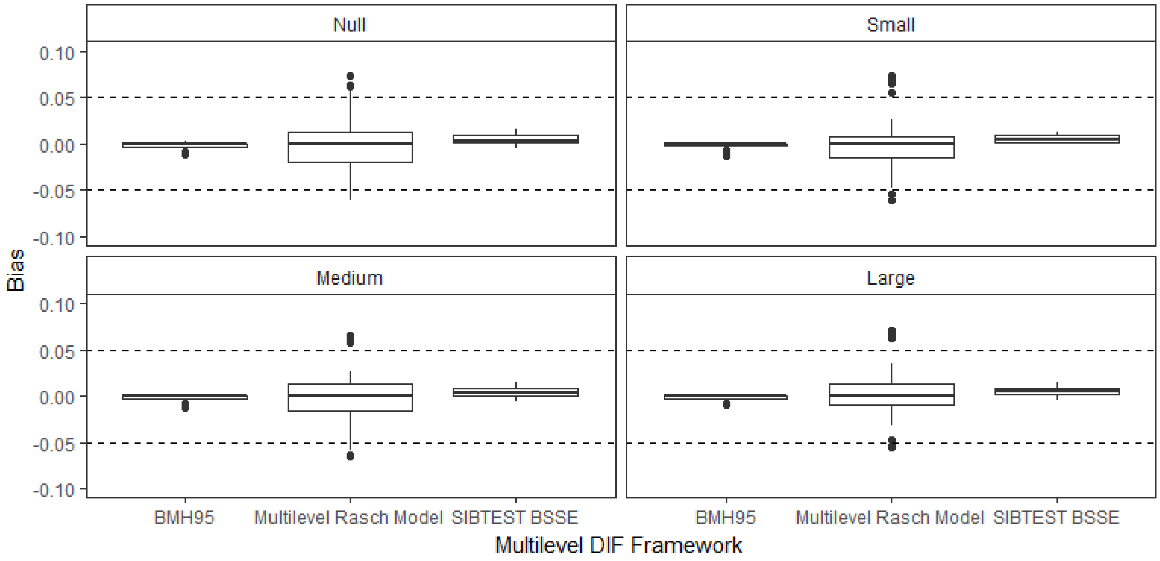

Figure 13. Relative Bias by Magnitude of DIF and Framework.

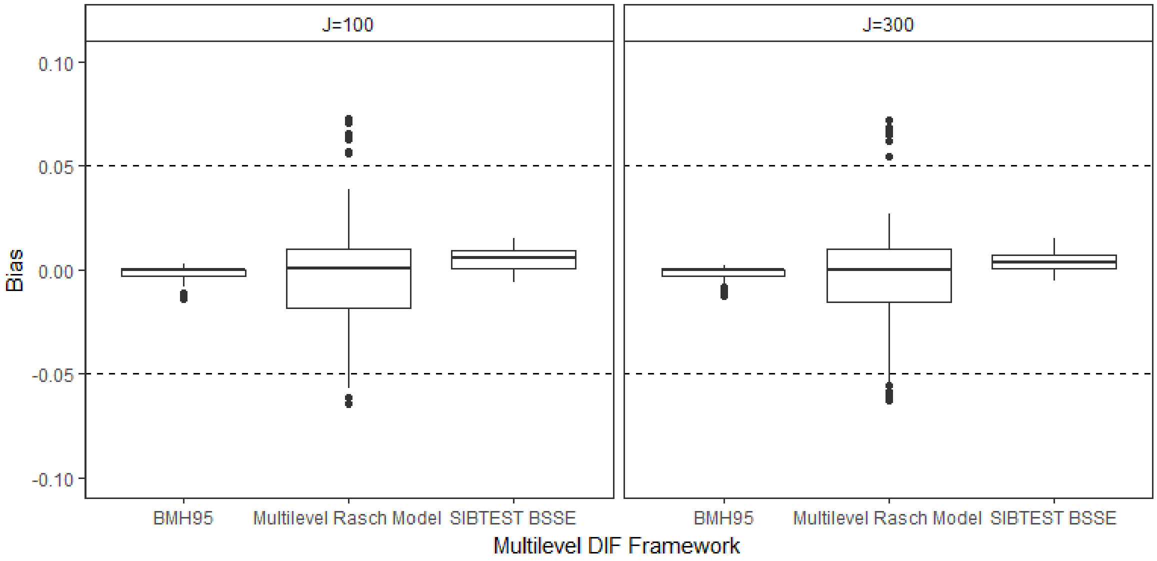

Figure 14. Relative Bias by Social-unit Sample Size and Framework.

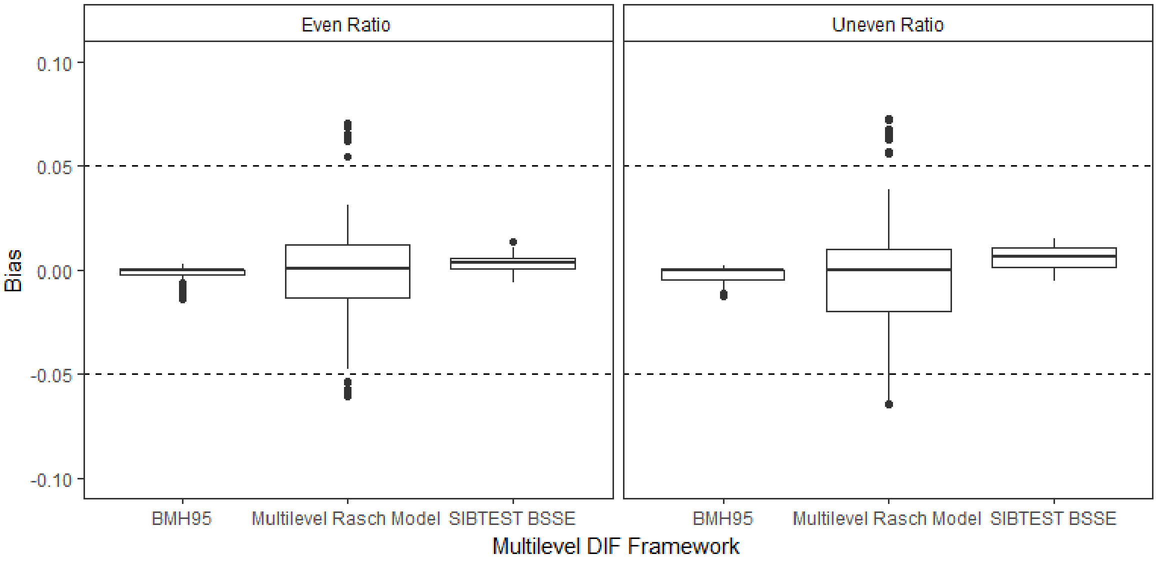

Figure 15. Relative Bias by Reference to Focal Group Ratio and Framework.

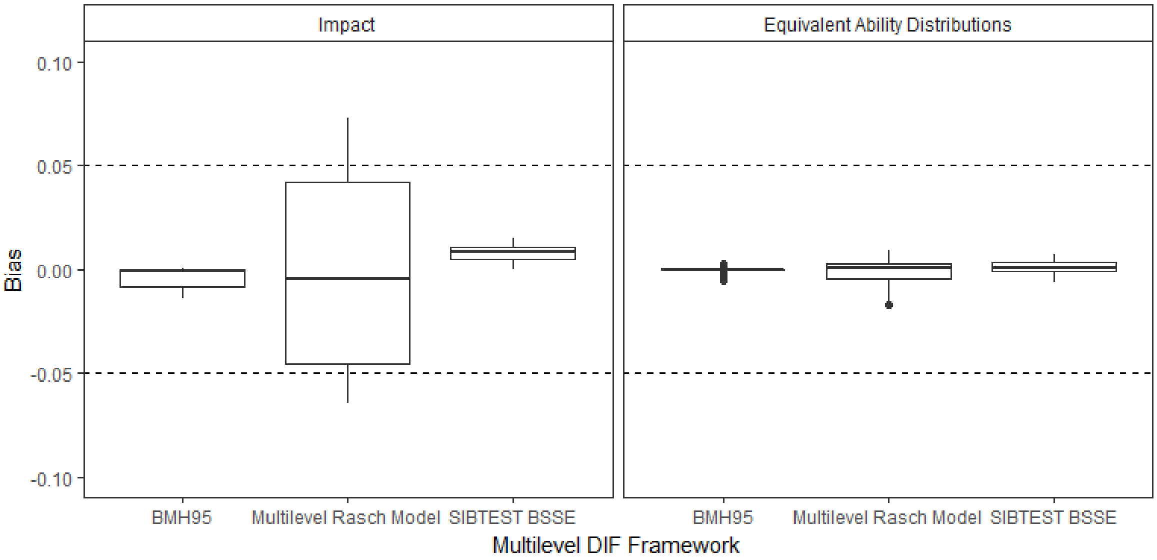

Figure 16. Relative Bias by Impact Presence and Framework.

A full factorial ANOVA was also used to determine which of the manipulated variables and interactions significantly affected the bias. Full results are presented in Appendix B, Table B.7.   In addition to the simulation conditions, the type of item was included as a factor to determine if the patterns of bias were different for invariant versus DIF containing items. A second full factorial ANOVA was run which introduced which group the item favored as a factor. A comparison of the two ANOVA models indicated that the inclusion of which group the item favored as a favor did lead to a significantly improved fit (p < 0.001). Results for the more complex model will be discussed subsequently.

There was a moderately large and significant interaction between which group the item favored and the multilevel DIF framework $F_{(2,6896)}$=12711.77, p<0.001, η^2=0.209. Figure 17 depicts the relationship between the favored group and the multilevel DIF framework. Across items which favored the focal and reference group and those which were invariant, the multilevel Rasch model had relatively similar bias. When the item was invariant, 24% of replications had relative bias outside of acceptable limits across all other conditions. When the item favored the focal group, 34% of replications had relative bias outside of the acceptable limits, all of which were underestimating the effect size. When the item favored the reference group, 16% of replications had relative bias outside of the acceptable limits, divided between underestimating and overestimating the effect size.

Conversely, the SIBTEST BSSE and BMH95 tended to overestimate the effect size for items which favored the focal group and underestimate the effect size for items which favored the reference group. For SIBTEST BSSE, all replications were outside of the acceptable limits for relative bias for items which favored the focal or reference groups. The BMH95 also suffered from extreme relative bias for items which favored the focal or reference groups. All replications were outside of the acceptable limits.

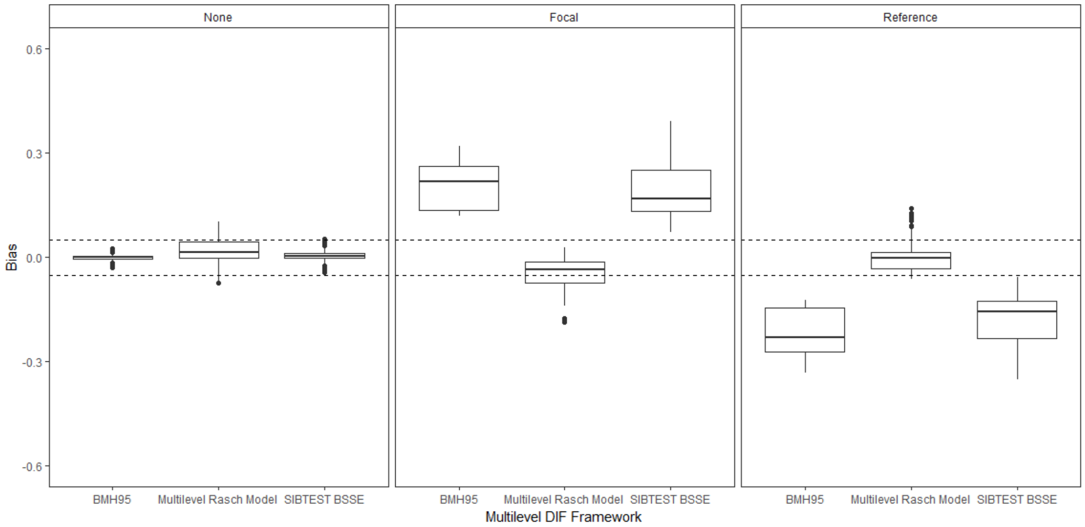

Figure 17. Interaction between Group Favored by DIF Item and Multilevel DIF Framework.

However, since the purpose of this study is primarily a comparative analysis between the three multilevel DIF frameworks the results from the simpler ANOVA model will be presented as well. When the favored group factor was removed, the multilevel DIF framework condition appeared in significant and practically large interactions. The highest order significant interaction was the three-way interaction between item type, the equivalency of the ability distributions and the multilevel DIF framework, $F_{(2,192)}$=1481.17, p<0.001, η^2=0.24. Figure 18 depicts the relationship between item type, impact status, and the multilevel DIF framework. The mean and standard deviation for these conditions is presented in Table 26.

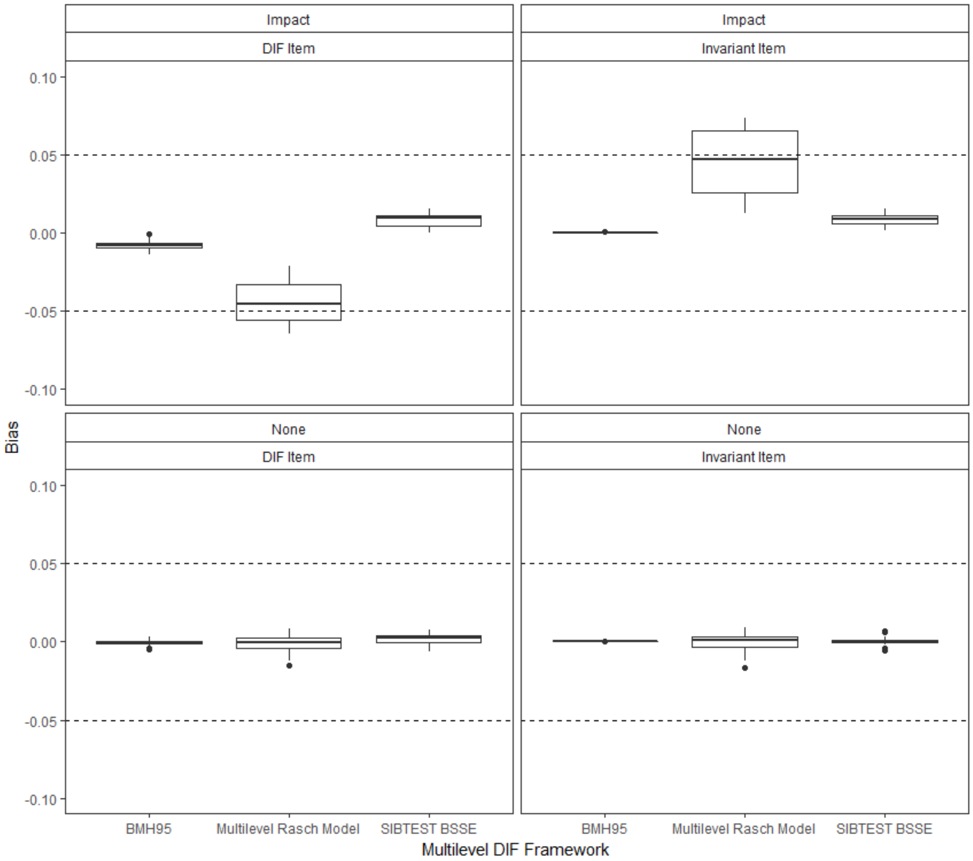

Figure 18. Bias by Item Type, Impact Presence, and Multilevel DIF Framework.

Table 26

Mean and Standard Deviation of Bias by Item Type, Impact Presence, and Multilevel DIF Framework

	Equal Ability Distributions		Impact	
	Invariant Item	DIF Item	Invariant Item	DIF Item
BMH95	0.0000 (0.000)	-0.0005 (0.002)	0.0000 (0.000)	-0.0079 (0.003)
Multilevel Rasch Model	-0.0003 (0.006)	-0.0016 (0.006)	0.0455 (0.020)	-0.0446 (0.012)
SIBTEST BSSE	0.0003 (0.002)	0.0021 (0.003)	0.0084 (0.004)	0.0083 (0.004)

All of the multilevel DIF frameworks resulted in relatively unbiased effect size estimates when the ability distributions of the focal and reference group were equivalent. However, under the impact condition the multilevel Rasch framework resulted in consistently biased estimates. The effect size estimates for the invariant items tended to be slightly over estimated while the effect size estimates for the items containing DIF tended to be slightly under estimated. For invariant items under the impact condition, the relative bias was outside of the acceptable range approximately 50% of the time. When impact was present, the relative bias was outside of the acceptable range approximately 34% of the time for DIF containing items.

Figure 19 portrays the RMSE across item type, impact status, and multilevel DIF framework. This interaction when predicting RMSE is both insignificant and has a negligible effect size, $F_{(2,192)}$=.030, p=0.970, η^2=0.000. However, a visual representation of the RMSE juxtaposed with the relative bias is helpful for understanding the effect size estimates, particularly those obtained by the BMH95 and SIBTEST BSSE. Although the bias results presented above indicate that the BMH95 and SIBTEST BSSE methods maintain acceptable relative bias rates the RMSE results demonstrate that this is due to highly variable data which is merely unbiased in a specific direction. While the multilevel Rasch model suffers from bias it has superior precision under these conditions.

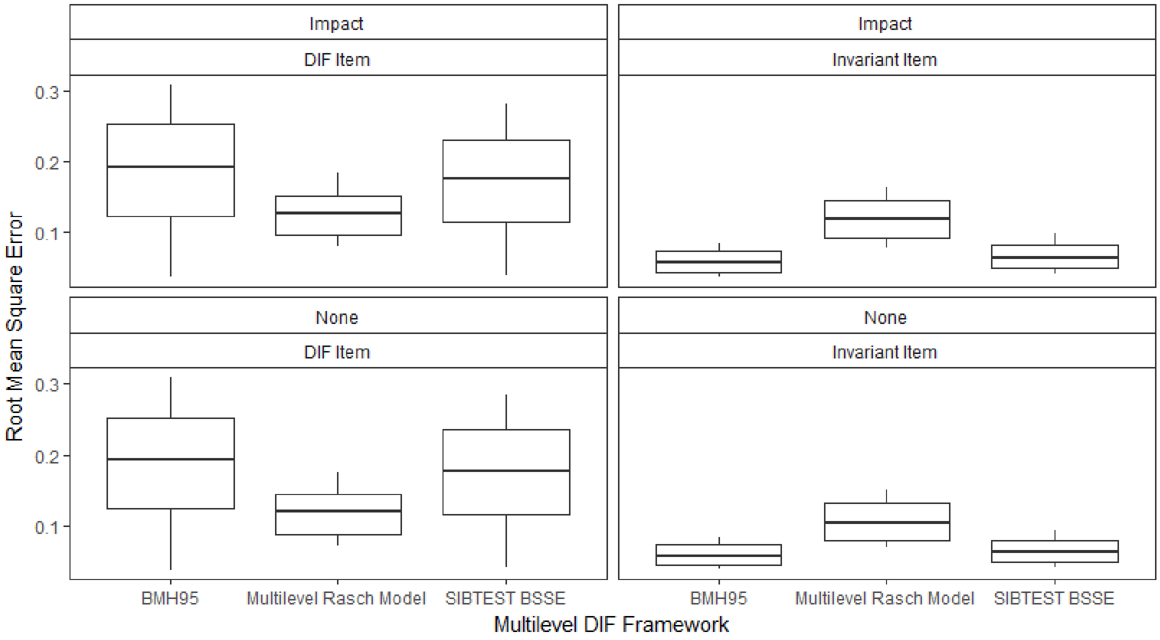

Figure 19. Root Mean Square Error by Item Type, Impact Status, and Multilevel DIF Framework.

Root Mean Square Error. The RMSE estimates for the effect size are presented in Figures 20 through 24. The RMSE estimates across conditions are presented in C1 through C6 in Appendix C. Lower RMSE values indicate better estimates of the effect size.

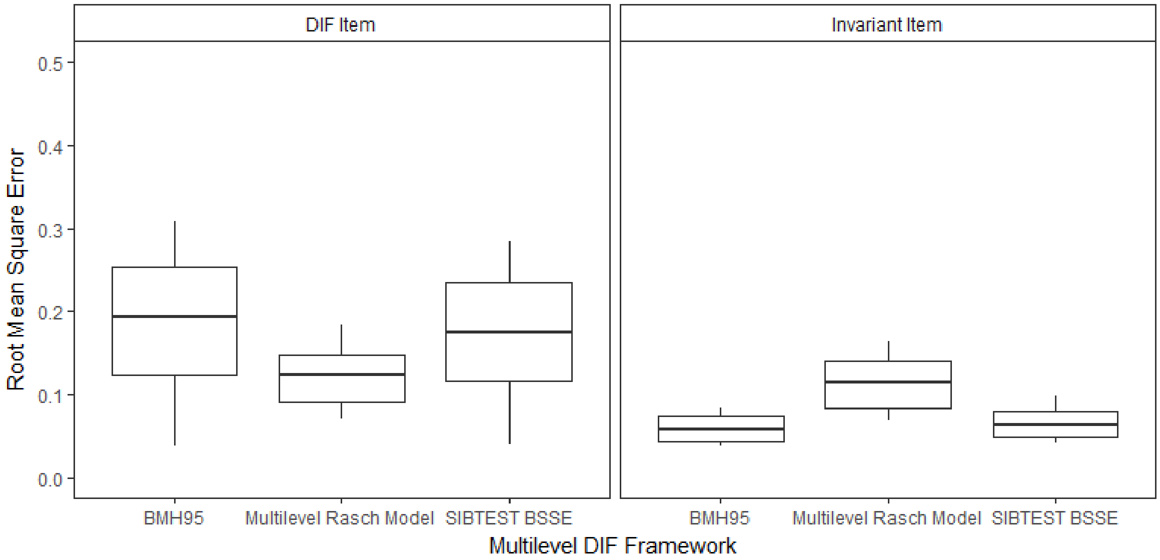

Figure 20. Root Mean Square Error by Item Type and Multilevel DIF Framework.

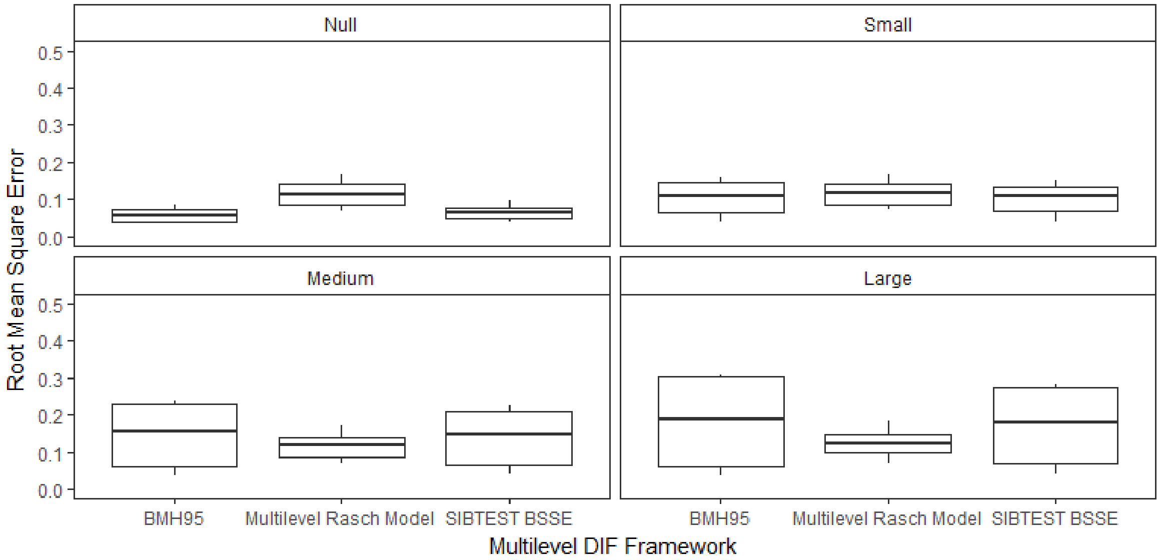

Figure 21. Root Mean Square Error by DIF Magnitude and Multilevel DIF Framework.

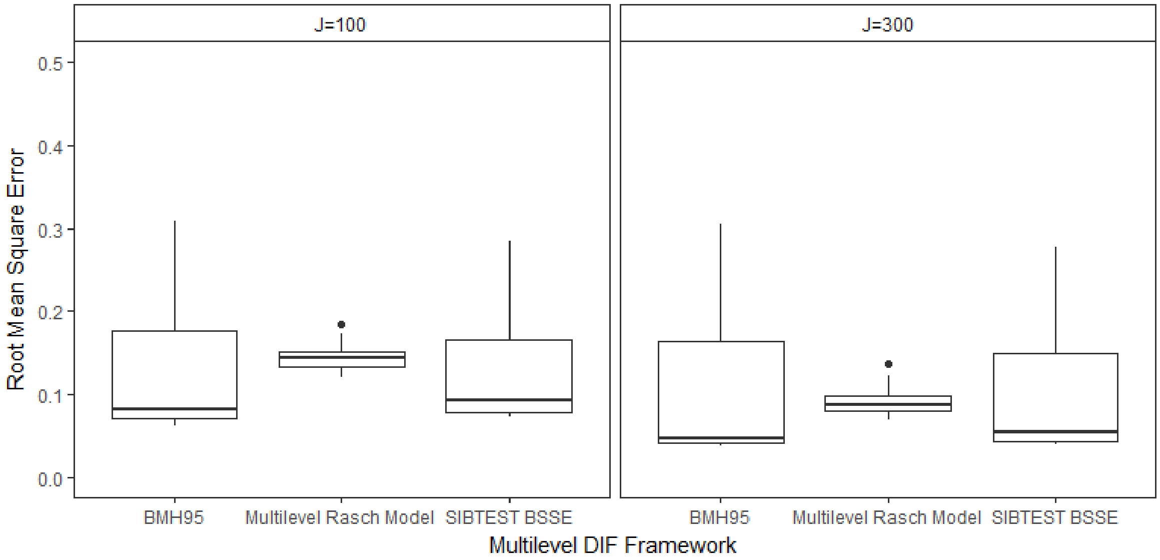

Figure 22. Root Mean Square Error by Social-unit Sample Size and Multilevel DIF Framework.

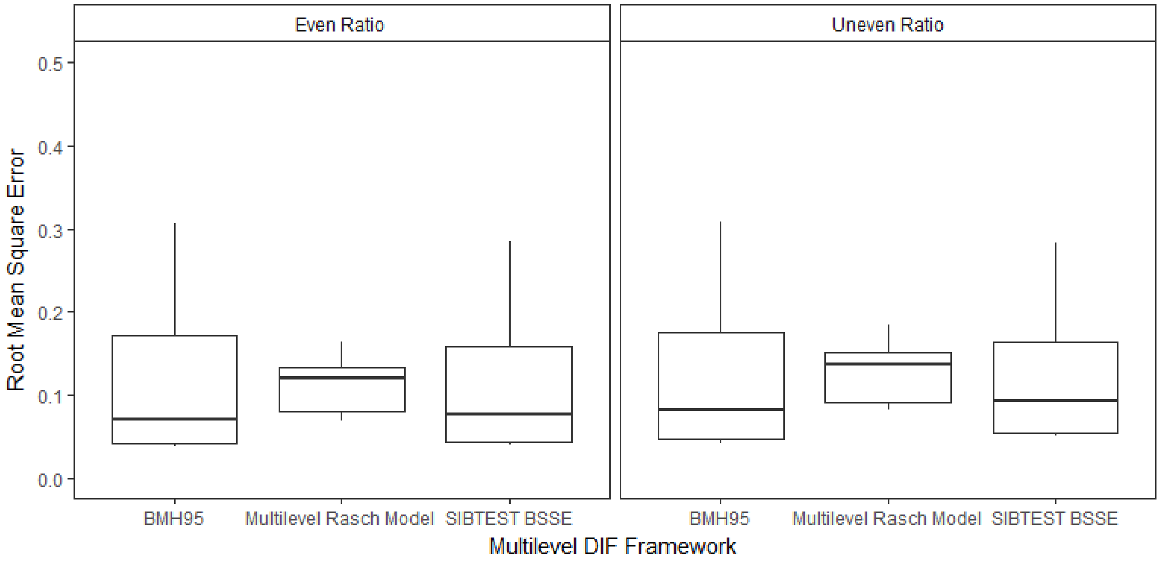

Figure 23. Root Mean Square Error by Reference to Focal Group Ratio and Multilevel DIF Framework.

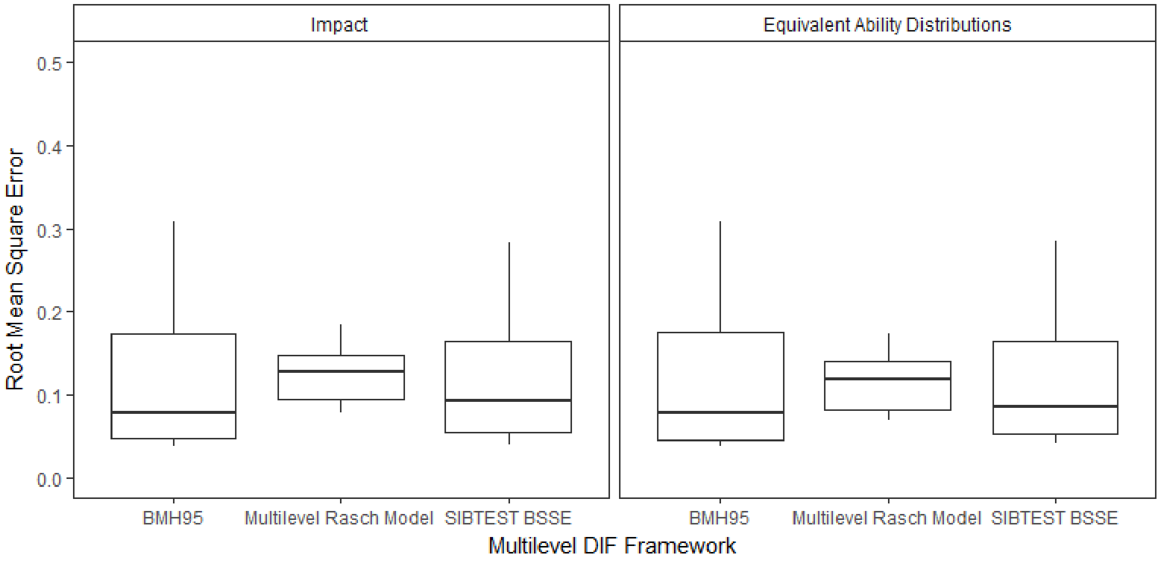

Figure 24. Root Mean Square Error by Impact Presence and Multilevel DIF Framework.

A full factorial ANOVA was also used to determine which of the manipulated variables and interactions significantly affected the RMSE. A second full factorial ANOVA was run which introduced which group the item favored as a factor. A comparison of the two ANOVA models indicated that the inclusion of which group the item favored as a favor did lead to a significantly improved fit ($p < 0.001$). However, in the second full factorial ANOVA, only the main effects of size and multilevel framework were significant with effect sizes of η^2 greater than 0.01. Therefore, the results for both models will be presented as the significant and practically large higher-order interactions in the simpler model add valuable information regarding the comparison of the three multilevel DIF frameworks. The simpler model will be presented first.

The highest order significant interaction was the three-way interaction between item type, invariant or containing DIF, magnitude of DIF and multilevel DIF framework. The interaction between item type, invariant or containing DIF, by magnitude of DIF by multilevel DIF framework was significant and moderately large, $F_{(3,256)}$=30.21, p<0.001, η^2=0.070. The interaction will be presented in two parts, first the general trends observed for the interaction of item type by magnitude of DIF and then the larger interaction including multilevel DIF framework. Figure 25 includes the RMSE for the magnitude of DIF by item type. As would be expected, under the null DIF condition items which were held invariant and those which were manipulated under DIF conditions had comparable RMSE levels. Across all other conditions, as the magnitude of DIF increased the average RMSE of invariant items remained stable. Indicating that even under large magnitude DIF conditions, the multilevel DIF frameworks produced effect size estimates for the invariant items that were comparable to those produced under the null condition. However, as the magnitude of DIF increased the average RMSE for the DIF containing items increased as well as did the distribution of observed RMSE values. Indicating that the effect size estimates were more precisely estimated in small magnitude DIF conditions.

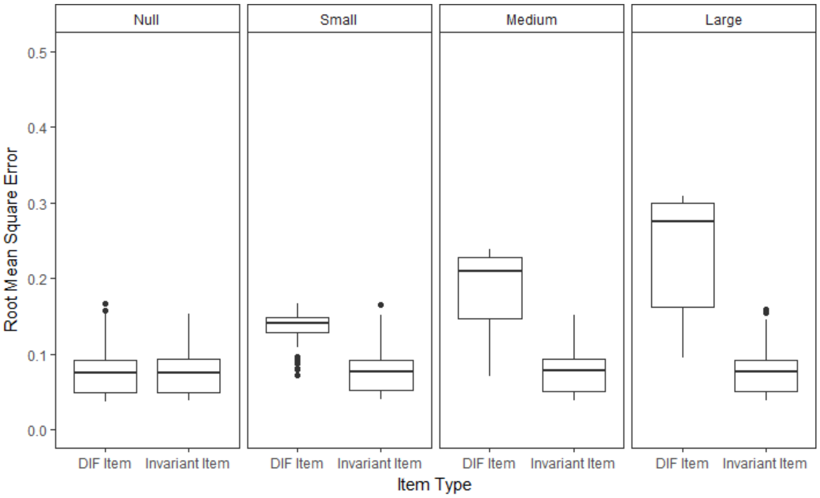

Figure 25. Root Mean Square Error by Item Type and Magnitude of DIF.

Figure 26 depicts the RMSE by item type, magnitude of DIF, and multilevel DIF framework. As depicted in Figure 26, across all other conditions the BMH95 and SIBTEST BSSE have comparable average RMSE for invariant items. The multilevel Rasch model less precisely estimates effect sizes, which should be zero, for invariant items across all other conditions. However, the multilevel Rasch model produces much more precise and accurate effect size estimates for items containing DIF. This is particularly true under the medium and large magnitude DIF conditions. This result is to be expected as the multilevel Rasch model is parameterized to provide effect size estimates for social-unit DIF, while the SIBTEST BSSE and BMH95 modifications are not designed to adjust effect size estimates.

The main effect of social-unit sample size was also significant, $F_{(1,256)}$=.116, p<0.001, η^2=0.058. The observed RMSE was higher when the number of clusters was 100 versus 300. Figure 27 depicts the relationship between social-unit sample size and RMSE.

In the second ANOVA, only the main effects of size and multilevel DIF framework were significant with large effect sizes. These main effects were also significant in the first ANOVA with small effect sizes. Figure 28 depict the relationships between RMSE and the multilevel DIF framework. Across all other conditions, the multilevel Rasch model had higher average RMSE. However, the multilevel Rasch model had a smaller range in RMSE values than the BMH95 and SIBTEST BSSE multilevel DIF frameworks.

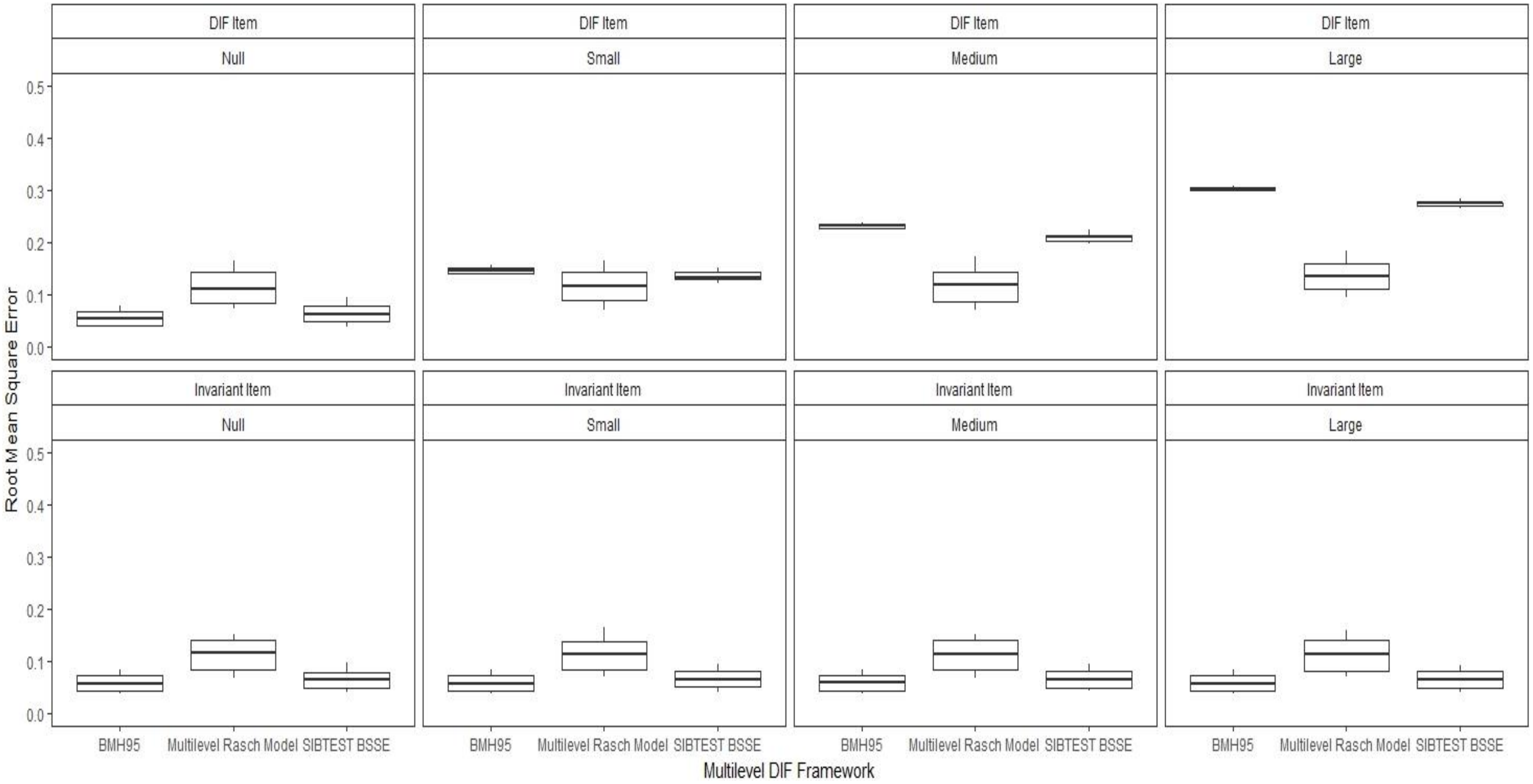

Figure 26. Root Mean Square Error by Multilevel DIF Framework, Item Type, and Magnitude.

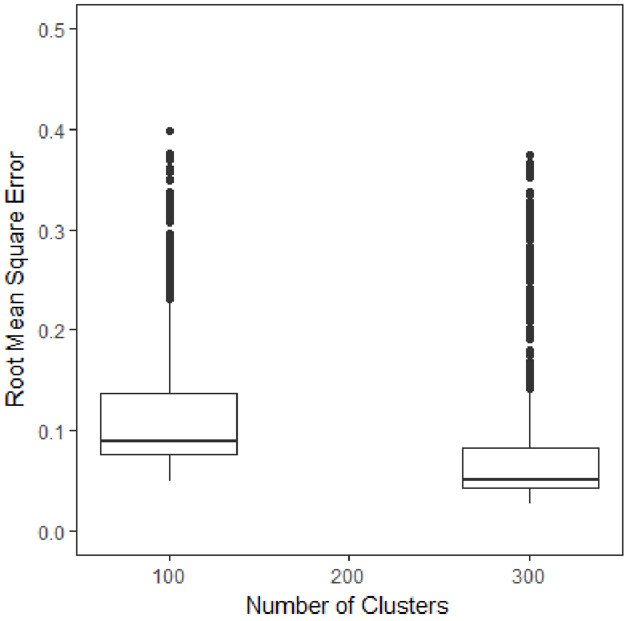

Figure 27. Root Mean Square Error by Social-unit Sample Size.

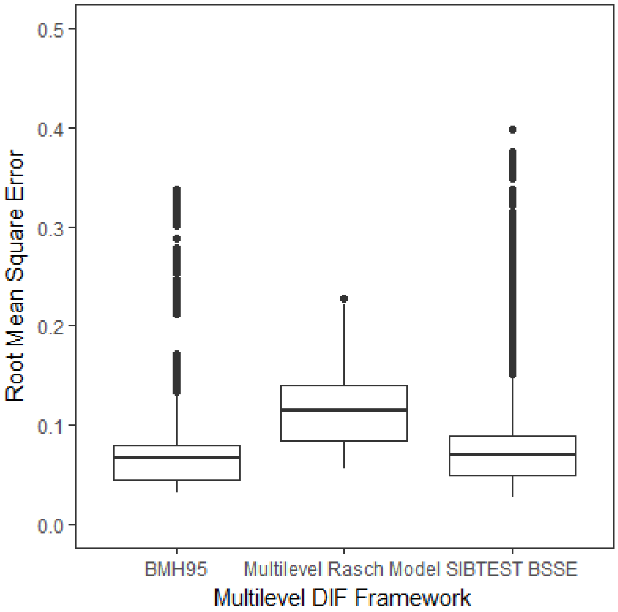

Figure 28. Root Mean Square Error by Multilevel DIF Framework.

Impact

Of the three multilevel DIF frameworks studied, only the multilevel Rasch model provides estimates for the difference between ability distributions of the focal and reference group. Although this cannot be a comparative analysis, since the data was captured it is instructive to understand the accuracy of impact estimates produced by the multilevel Rasch model under various conditions. Particularly, as there was scant published research investigating multilevel DIF detection utilizing a three-level model in the presence of impact. For that reason, mean impact estimates, relative bias and RMSE shall be presented.

The mean and standard deviation for the impact estimates are provided in Table 27. Additional salient information includes the correct identification of the reference group as the group with the higher mean ability for the dimension of interest. In general, the reference group was accurately selected as the group with the higher mean ability, with a range of 93%-100% of replications. When the social-unit sample size was *J*=100 values fell below 100%, while when *J*=300 the reference group had the higher mean ability across 100% of replications. When the ability of both groups were equivalent, the reference group was identified as having a higher mean ability in 41%-62% of replications. Results are expected as neither group had a higher mean ability and indicates that in general the model didn’t favor one group over the other.

Table 27

Impact Analysis

		High (ICC = 0.2)				Low (ICC = 0.1)			
		None (Δ=0.0)		Impact (Δ=0.5)		None (Δ=0.0)		Impact (Δ=0.5)	
		Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)
Null									
(δ=0.0)	100	0.20 (0.17)	0.23 (0.18)	0.48 (0.24)	0.51 (0.28)	0.10 (0.07)	0.10 (0.08)	0.45 (0.13)	0.43 (0.15)
	300	0.12 (0.08)	0.13 (0.09)	0.46 (0.15)	0.44 (0.20)	0.06 (0.04)	0.08 (0.05)	0.42 (0.08)	0.42 (0.09)
Small									
(δ=0.2)	100	0.18 (0.13)	0.22 (0.18)	0.45 (0.21)	0.43 (0.23)	0.09 (0.07)	0.11 (0.08)	0.43 (0.13)	0.42 (0.15)
	300	0.12 (0.09)	0.11 (0.08)	0.47 (0.15)	0.46 (0.15)	0.05 (0.04)	0.07 (0.05)	0.41 (0.07)	0.40 (0.08)
Medium									
(δ=0.4)	100	0.19 (0.15)	0.21 (0.17)	0.48 (0.21)	0.52 (0.27)	0.10 (0.08)	0.11 (0.08)	0.41 (0.13)	0.42 (0.14)
	300	0.10 (0.07)	0.12 (0.09)	0.45 (0.14)	0.49 (0.16)	0.06 (0.05)	0.07 (0.05)	0.41 (0.08)	0.41 (0.08)
Large									
(δ=0.6)	100	0.18 (0.14)	0.21 (0.15)	0.47 (0.20)	0.50 (0.27)	0.10 (0.07)	012 (0.09)	0.40 (0.11)	0.41 (0.15)
	300	0.39 (0.08)	0.12 (0.09)	0.46 (0.14)	0.45 (0.17)	0.06 (0.05)	0.07 (0.05)	0.39 (0.08)	0.41 (0.09)

Relative Bias. A full factorial ANOVA was also used to determine which of the manipulated variables and interactions significantly affected the bias. The bias estimates across conditions are presented in D1 in Appendix D. The highest-order significant interaction with at least a small effect size ($\eta^2 > 0.01$) was the interaction social-unit sample size by impact presence, $F_{(1,3)}$=55.53, *p*=0.005, η^2=0.015. Figure 29 depicts the relationship between impact presence and social-unit sample size.

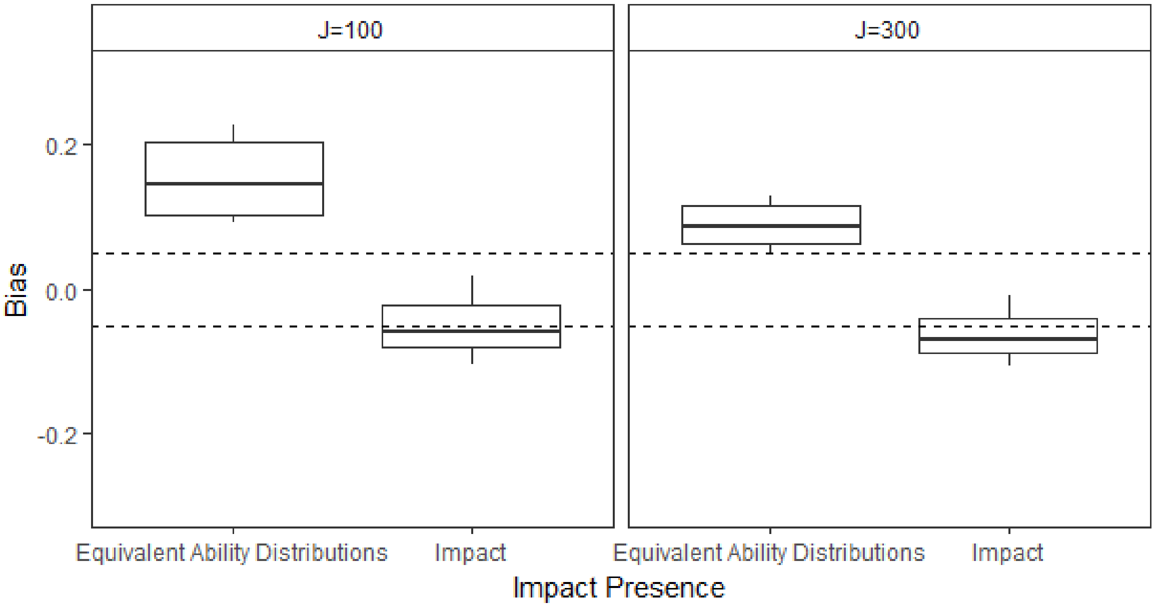

Figure 29. Bias by Impact Presence and Social-unit Cluster Size.

When impact was present, the difference in ability distributions between the reference and focal group tended to be underestimated (*M*=-0.06, *SD*=0.03). The bias estimates were outside the acceptable range 53% of the time, all underestimated. On average, estimates were more severely underestimated with larger number of clusters (*M*=-0.06, *SD*=0.03) than with a smaller number of clusters (*M*=-0.05, *SD*=0.04). When the ability distributions were equal the difference in ability distributions between the reference and focal group, which should be zero, were overestimated (*M*=0.12, *SD*=0.05). One hundred percent of bias estimates were outside of the acceptable range, all overestimated. On average, estimates were more severely overestimated with a smaller number of clusters (*M*=0.15, *SD*=0.05) than with a larger number of clusters (*M*=0.09, *SD*=0.03).

The main effect of the magnitude of ICC was also significant, $F_{(1,3)}$=39.51, p<0.001, η^2=0.11. Under the high magnitude ICC condition, estimates of impact (*M*=0.06, *SD*=0.10) were slightly more prone to overestimation than under the low magnitude ICC condition (*M*=0.00, *SD*=0.09). Fifty-six percent of bias estimates under the high magnitude ICC condition were outside the acceptable range, 97% of bias estimates were outside the acceptable range under the low magnitude ICC condition. Although the interaction was not statistically significant, impact estimates under the low magnitude ICC and impact condition were underestimated while those under the low magnitude ICC and equivalent ability distribution condition were overestimated.

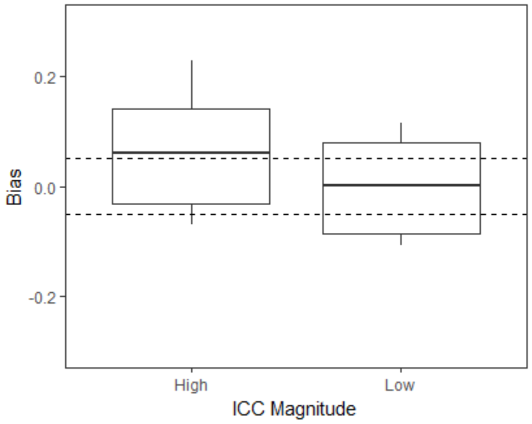

Figure 30. Bias Across ICC Magnitudes.

Root Mean Square Error. A full factorial ANOVA was also used to determine which of the manipulated variables and interactions significantly affected the root mean square error. The highest-order significant interactions with at least small effect sizes ($\eta^2 > 0.01$) were the interaction of social-unit sample size by impact presence, the interaction of social-unit sample size by ICC magnitude, and the interaction of impact presence by ICC magnitude. The results will be presented in order of decreasing effect size magnitude.

The highest order significant interaction with a medium effect size was the interaction between social-unit sample size and magnitude of ICC, $F_{(1,3)}$=.0117, p<0.001, η^2=0.05. In general, under the high magnitude ICC condition impact estimates were less precise than under the low magnitude ICC condition. This is particularly true when the number of clusters equaled 100.

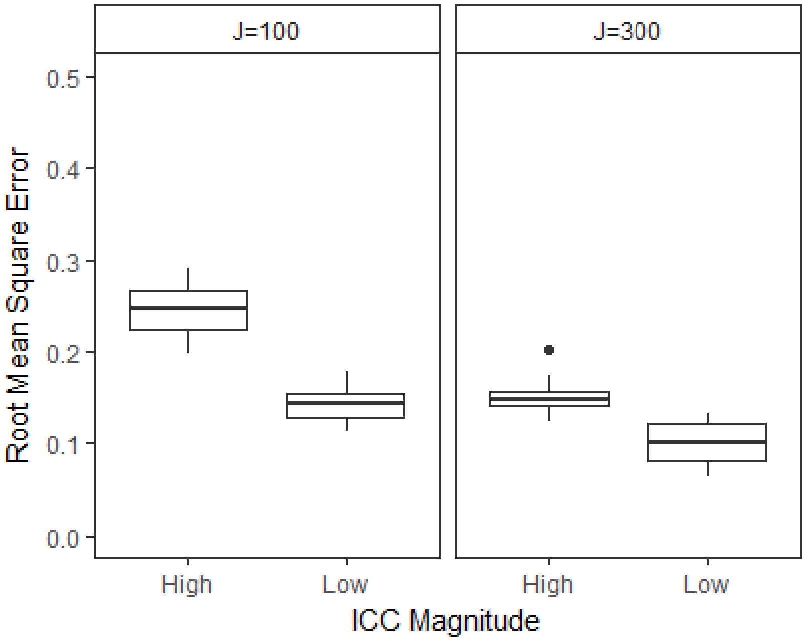

Figure 31. Root Mean Square Error by ICC Magnitude and Social-unit Sample Sizes.

The interaction between impact presence and ICC magnitude was also significant and had a small effect size, $F_{(1,3)}$=40.50, *p*=0.008, η^2=0.025. In general, estimates were more precise under the low ICC magnitude than the high. Estimates were also more precise when impact was present rather than when the ability distributions of the two groups were equivalent.

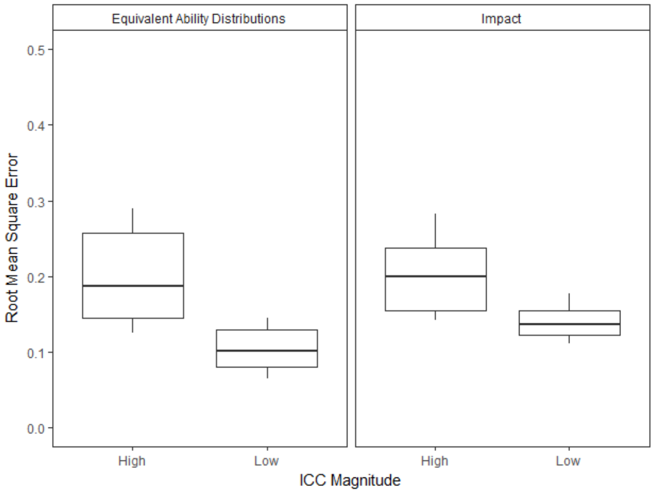

Figure 32. Root Mean Square Error by ICC Magnitude and Impact Presence.

Lastly, the interaction between social-unit sample size and impact presence was also significant and had a small effect size, $F_{(1,3)}$=17.54, *p*=0.025, η^2=0.011. In general, under the condition of larger social-unit sample size, estimates were more precise than under the condition of smaller social-unit sample size. However, this is particularly true when impact was present. Impact estimates were less precise when the ability distributions between the two groups were even. This trend is more pronounced in the condition of larger social-unit sample size.

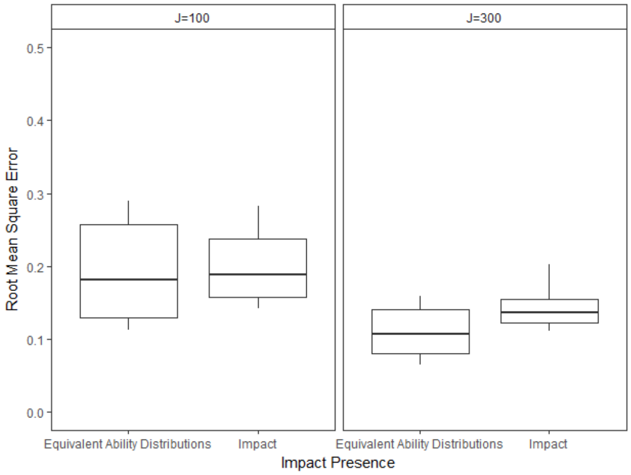

Figure 33. Root Mean Square Error by Impact Presence and Social-unit Sample Size.

Impact Discussion. Comparing the relative bias and RMSE results indicates that in general the multilevel Rasch model had superior performance when estimating impact rather than estimating equivalent ability distributions between the two groups. Additional factors that significantly improved performance were larger number of social-unit clusters and lower magnitude ICC.

Validation Implications

The above results will be discussed first in terms of the five sources of evidence from *The Standards* (2014). Then the appropriateness of each multilevel framework will be discussed first in relation to the multilevel IUA and then in relation to an ecological model of validation.

Regardless of the multilevel validation framework chosen, the statistical results indicate that there are indeed methodologies capable of supporting examinations of DIF at the social-unit level. Therefore, results will be presented as a discussion of the types of scenarios which may arise from testing for social-unit DIF. Issues of evaluation will also be discussed. Across sources, care would need to be taken when universally using any multilevel DIF framework, as the statistical results do not support such a use. For example, in cases when impact is present the multilevel Rasch model would be less appropriate than the BMH95.

Sources of Evidence at the Social-unit Level. The discussion below assumes social-unit main effect DIF, as was simulated in this study. Alternative interpretations of social-unit DIF would exist in the presence of main effect DIF, interactional DIF, or examinee level variable DIF, more complex studies than what was undertaken here. For instance, with examinee level variable DIF a social-unit grouping covariate can be introduced to explain variation in examinee level DIF that occurs across social-units. In this case, the social-unit grouping covariate can be viewed as a moderator rather than a social-unit DIF causing characteristic.

Test Content. DIF which arises at the social-unit level may have direct implications for understanding the accessibility of test content. As an example, DIF analyses could investigate the differential effects of school curriculums on item difficultly. Statistical differences would not necessarily equate to bias but may indicate that some curriculums are better aligned to test content. If it was determined that the test content was representative of the construct it was intended to measure this information could have administrative implications but would not likely result in items being removed from the test.

Conversely, if the DIF analyses focused on the differential effects of school poverty on item difficultly results may have more of a traditional DIF outcome. If school poverty yielded statistical DIF for some items then the ultimate conclusion may be reached that students at a particular type of school have less access to the content than others. While this may have administrative implications, this scenario aligns more closely with traditional examinee level DIF investigations and would likely have similar outcomes (e.g. bias panels and potential item removal).

Response Processes. Identification of social-unit DIF indicates that response processes are not only shaped by individual characteristics but by social-unit characteristics as well. The results of the current study would support the assertion that response processes are only differentially affected by social-unit characteristics. However, more complex types of DIF can be studied which begin to address the issue of interaction between item, person, and societal characteristics on response processes.

Internal Structure. The identification of social-unit DIF speaks directly to issues related to internal structure. Investigating social-unit DIF broadens DIF analyses from focusing on groups identified by individual characteristics to groups which are defined by social-unit characteristics, such as school poverty or teaching style. Evidence of social-unit DIF as a product of construct-irrelevant variance would detract from the validity of test score interpretation and use. Interpretation would likely be similar to examinee level DIF analyses.

Relations to Other Variables. Social-unit DIF would imply that the relationship between item responses and external criterions is not the same across groups. Using predictive validity as an example, if the social-unit DIF was appreciably large and favored one group it could result in the measure not being an adequate predictor of another future measure. This is particularly troubling given the nature of accountability testing and the focus on predicting which students will succeed academically and professionally.

From a convergent validity standpoint, it could raise questions of face validity if two measures had low correlation due to social-unit DIF in one measure. For instance, consider social-unit DIF occurring at the school level. Teachers and administrators may be less inclined to trust the results of the assessment, parents may feel the assessment is a waste of their children's time, and general disenfranchisement with the measure could set in.

Consequences of Testing. If DIF is found at the social-unit level, then an evaluation of intended positive and unintended negative consequences related to test use should be undertaken. As is the case when investigating examinee level DIF, it should be determined that the positive outcomes outweigh all unintended negative consequences. Such an analysis, coupled with measures of effect size, would provide direction regarding if an item was to be maintained or eliminated from an item bank.

Multilevel Interpretation Use Argument. The statistical results indicate that in general the three multilevel DIF frameworks adequately detected DIF while maintaining the Type I error rate. The study design is most similar to the two-pronged approach presented as a multilevel adaptation of an IUA. Social-unit DIF was tested in the absence of main effect person DIF, interactional DIF, or examinee level variable DIF. If the inferences at Level 1 and Level 2 of the multilevel IUA are to be considered independently then this is likely the type of analyses that would be undertaken.

In regards to tests of statistical significance, the BMH95 had superior results. However, the multilevel Rasch model produced less systematically biased and more precise estimates of effect size for items containing DIF. The appropriateness of a multilevel DIF framework would likely be determined by a practitioner's needs. All of the studied multilevel DIF frameworks are appropriate in nature to provide evidence for the multilevel IUA as proposed in this study. However, their efficacy at detecting and quantifying DIF is dependent upon simulation conditions.

Ecological Model of Validation. The statistical results indicate that in general the three multilevel DIF frameworks detected DIF while maintaining the Type I error rate. However, the BMH95 and SIBTEST BSSE frameworks are more appropriate for a consideration of DIF under a multilevel IUA. Specifically, they consider the nesting of data but do not consider the interaction that occurs between those levels of nesting. A more appropriate visual for validation under the ecological model when evidence is provided by the BMH95 and SIBTEST BSSE methods is provided in Figure 34.

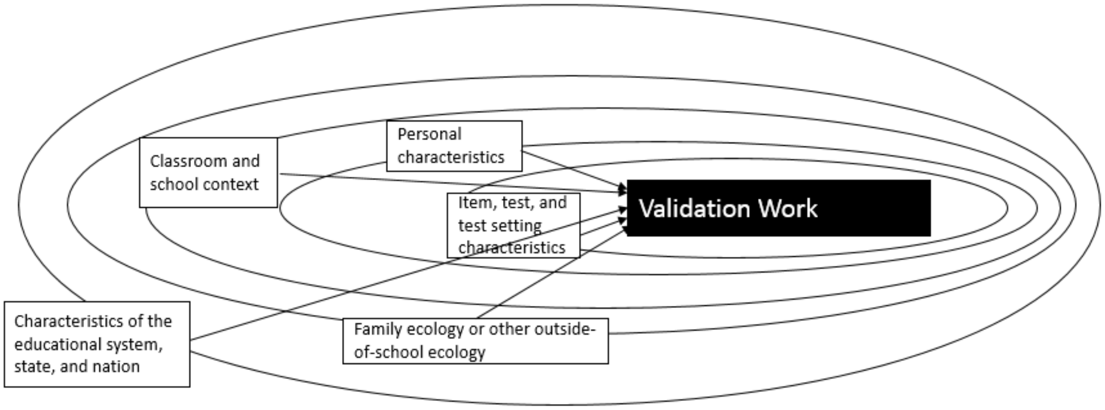
Personal characteristics
Classroom and school context
Validation Work
Item, test, and test setting characteristics
Characteristics of the educational system, state, and nation
Family ecology or other outside-of-school ecology

Figure 34. A Non-Interactional Ecological Model.

In this version the porous lines which separated each level of nesting have been replaced by solid lines. Thus, demonstrating that while the data is indeed nested, information is not being exchanged across levels. Likely validation under this model would proceed similarly to the adapted multilevel IUA. Information from each level would be collected separately and validation work would continue for each level of nesting independent of all other levels. While failing to validate the test score interpretations and use at a lower level may raise red flags for the interpretation and use of test scores at a higher level, interaction between the two levels would not exist.

In such a context, the empirical evidence gathered in this study suggests that the BMH95 and SIBTEST BSSE will provide adequate evidence for the detection of social-unit DIF. However, due to higher power rates in the presence of small magnitude DIF the BMH95 would be the preferable framework.

Conversely, the multilevel Rasch model allows for interaction between levels of nesting as well as the incorporation of item level characteristics. However, the results of this study only support the use of the multilevel Rasch model in a main effect scenario. While the multilevel Rasch model is capable of supporting more advanced frameworks, this study’s results support a multilevel ecological model such as the one presented in Figure 34.

If a researcher is espousing a *Third Generation DIF* mentality then the multilevel Rasch model provides more robust information than the BMH95 or SIBTEST BSSE. Of the three DIF frameworks presented it is the only method capable of incorporating item features into DIF analysis. If a researcher aims to explore more complex DIF scenarios, the multilevel Rasch model is also the only appropriate multilevel DIF framework of the three presented.

CHAPTER V

DISCUSSION

The purpose of the current study was twofold. First, to highlight the need for multilevel validation frameworks considering our education policy focus on accountability and the naturally nested structure of education data. Within this discussion, multilevel DIF frameworks were identified as a promising method for providing validation evidence. The second goal was to examine the impact of various conditions on the Type I error and power rates of multilevel DIF frameworks and the estimated effect sizes. The theoretical considerations will be discussed before moving onto the statistical results.

Multilevel Validation and DIF

Researchers have presented ample challenges to current validity theories and validation practices. The main challenges, as relevant to the discussion of DIF and accountability, include: the assertion that context and item/test performance are inextricably linked (Chalhoub-Deville, 2003; Zumbo et al., 2015), the increasing demand for the evaluation of actual and intended score use and interpretation (Bennett et al., 2011; Hubley & Zumbo; Moss, 2016; Sireci, 2016), and where we situate the evidence provided by DIF analyses in the validation process (Gomez-Benito, 2018; Walker, 2011).

These challenges are beginning to be addressed in the literature via revised validation frameworks and the championing of more advanced methodologies for addressing DIF. Specifically, researchers are addressing the multilevel nature of constructs and the need for validation at all levels of aggregated score interpretation and use (Chen et al., 2004; Forer & Zumbo, 2011). A comprehensive multilevel validation framework has yet to be presented. However, researchers are promoting the consideration of contextual factors when providing validation evidence for response processes (Chen & Zumbo, 2017; Zumbo et al., 2015). Others are demonstrating how existing validation frameworks can be adapted to the needs of aggregate score interpretation and use (Haertl, 2013).

The shift towards ecological modeling and validation at the social-unit level requires new methodologies for providing evidence. Reoccurring in the literature was the call for multilevel modeling to provide DIF evidence. However, in this new era of validation, DIF analyses would not only be used to flag items as potentially biased but to provide evidence regarding the root causes of bias. Solidly placing DIF analyses in Zumbo's (2009) *Third Generation*. While the primary focus of existing literature was on HGLM models for DIF analysis (Chen & Zumbo, 2017; Zumbo et al., 2015), additional methods are being presented (French & Finch, 2013; French & Finch, 2015). Not all methods will address the increasing demand for methodologies which help us understand DIF in addition to identifying it. However, all methods provide increased statistical rigor and are appropriate for providing validation evidence at the social-unit level.

Whether multilevel validation will become a widespread practice remains to be seen, however, it is promising that researchers have begun to outline concrete steps for multilevel validation (Chen et al., 2004) and have presented applied results (Chen &

Zumbo, 2017; Li et al., 2017; Zumbo & Forer, 2011). The recency of publication of the literature surveyed supports the notion that the concept of multilevel validation and the methodologies to support it are gaining traction. Of the twenty-four papers surveyed which centered around concepts of multilevel validation or multilevel DIF, only two were published prior to 2005. Future literature outlining frameworks for multilevel validation will further the operational appeal of such endeavors.

The results from the simulation study, which will be discussed subsequently, highlight a significant consideration when undertaking multilevel validation using multilevel DIF frameworks. Not all multilevel DIIF frameworks are universally appropriate for use with all multilevel validation frameworks. The results from this study in particular provide empirical evidence supporting the use of multilevel DIF frameworks for validation that does not consider the interaction of characteristics across levels.

Study Findings and Conclusions

Three multilevel DIF frameworks were used to detect DIF and estimate an effect size. The first used the Beggs adjustment with a correction of 0.95 to the Mantel-Haenszel method to account for multilevel data (BMH95). The second was the bootstrap standard error adjustment for SIBTEST (SIBTEST BSSE). Lastly, the multilevel Rasch model was used to detect DIF. Manipulated conditions for the study included the magnitude of the DIF, the presence of impact, the ratio of focal to reference group sample size, intraclass correlation values, and the social-unit sample size. Conditions which were held constant include the proportion of items with DIF, the proportion of items favoring each group (focal or reference), and the within cluster sample size.

First, a brief summary is presented of the findings for each of the three research questions. Followed by general implications for multilevel DIF practices.

Research Question 1: Power and Type I Error Rates. In general, the nominal Type I error rates were maintained across the three multilevel DIF frameworks. The SIBTEST BSSE method tended to have extremely low Type I error rates but was susceptible to low power under certain conditions as well. The multilevel Rasch model appeared to be most sensitive to the simulation conditions with inflated Type I error rates observed when impact was present, and the ICC was high (0.20).

Across conditions, generally high power was observed with the exception of small magnitude DIF and a cluster sample size of *J*=100. As in other research, additional factors which were found to decrease the power rate include the presence of impact, increased ICC magnitude, and an uneven ratio of reference to focal group members. Of the three methods compared, the SIBTEST BSSE was observed to have the lowest power rates. While this was typically above acceptable rates, this was not the case for the condition of small magnitude DIF. Particularly when an uneven reference to focal group ratio was implemented a dramatic decrease in power was observed.

Research Question 2: Factors Influencing the Power and Type I Error Rates. A full factorial ANOVA was modeled with the Type I error rate as the dependent variable and all of the conditions within the study as the independent variables. Type I error rates were significantly related to all of the conditions within the study. However, only the interactions between the social-unit sample size by the multilevel DIF framework and presence of impact by multilevel DIF framework were significant and meaningfully large. The inclusion of social-unit sample size is consistent with other published findings.

The findings from the ANOVA indicate that the multilevel DIF frameworks performed differentially under the various simulation conditions. Across all conditions, the SIBTEST BSSE and BMH95 methods maintained the nominal error rate indicating that these two adjustments adequately adjust their single level counterparts to account for multilevel data. However, while the BMH95 maintained consistent Type I error rates, as the number of social-unit clusters increased, the Type I error rate of the SIBTEST BSSE decreased dramatically to consistently negligible error rates. Similar findings were observed by French and Finch (2015). Conversely, the multilevel Rasch model suffered from slightly inflated Type I error rates under the condition of *J*=300.

The effect of impact was also differential across the multilevel DIF frameworks. While both the BMH95 and SIBTEST BSSE maintained the nominal error rate in the presence of DIF, use of the multilevel Rasch model resulted in inflated Type I error rates. Of the impact conditions which resulted in inflated Type I error rates for the multilevel Rasch model, they exclusively occurred under the higher ICC condition. Indicating that at least for the multilevel Rasch model, as the ICC increases the model becomes less able to differentiate impact from DIF.

Based on the ANOVA results for the power rate, the interaction between the magnitude of DIF, social-unit sample size, and multilevel DIF framework used had a statistically significant small effect on the power rate. Specifically, it was found that with a smaller number of social-unit clusters (*J*=100) the average power rate failed to meet acceptable levels for small magnitude DIF. With a larger number of clusters, all magnitudes of DIF were detected with adequate power. Further segmenting the data revealed that none of the models reached an acceptable average power rate under the small social-unit sample size (*J*=100) and small magnitude DIF (0.3) condition. However, under a larger number of social-unit clusters (*J*=300) and small magnitude DIF (0.3) condition, the BMH95 and multilevel Rasch model reached adequate power rates while the SIBTEST BSSE did not. Indicating that while the SIBTEST BSEE maintains acceptable Type I error rates across conditions, it is not adequately powered in conditions of small magnitude DIF. These results confirm prior findings for the SIBTEST BSSE from French and Finch (2015) when using a slightly higher magnitude of DIF (0.4).

Research Question 3: Effect Size. Effect size estimates were evaluated using relative bias and RMSE. When analyzing the relative bias using a simpler ANOVA model, the interaction between impact status, the type of item and the multilevel DIF framework was the highest-level interaction with a large effect size. The BMH95 and SIBTEST BSSE had acceptable relative bias rates regardless of type of item or the presence of impact. However, the multilevel Rasch model tended to underestimate the effect size estimates for items containing DIF and overestimate the effect size estimates for items which did not contain DIF in the presence of impact. These results align with the results of the Type I error rate analyses. Not only did the multilevel Rasch model tend to overestimate the DIF effect size in the presence of impact, the results were considered significant at a level above what would be expected by chance. Therefore, in the presence of impact, researchers may make incorrect conclusions regarding the magnitude of DIF when using the multilevel Rasch model.

While the BMH95 and SIBTEST BSSE maintained acceptable relative bias, their RMSE values were higher on average when compared to the multilevel Rasch model for items containing DIF. Indicating that while the multilevel Rasch model underestimated effect sizes for items containing DIF the estimates were more consistent than those obtained with the BMH95 and SIBTEST BSSE methods. As a result, practitioners using the latter two methods would estimate effect sizes that would be unpredictably different from truth. However, for invariant items, both the BMH95 and SIBTEST BSSE had superior average RMSE values when compared to the multilevel Rasch model. This was true in both the presence and absence of impact.

When modeling RMSE, the three-way interaction between type of item, magnitude of DIF, and the multilevel DIF framework was the highest order significant interaction with a large effect size. Across magnitudes of DIF, the BMH95 and SIBTEST BSSE had superior RMSE values to the multilevel Rasch model for invariant items. However, as the magnitude of DIF increased for items containing DIF the RMSE values increased for both the BMH95 and SIBTEST BSSE. The observed RMSE values were quite large compared to the multilevel Rasch model, indicating that across other conditions the effect size estimates using both non-parametric multilevel DIF frameworks were quite variable in comparison.

A secondary analysis was conducted which included which group an item favored, either reference or focal, as a factor. While the DIF was balanced between groups the impact condition always favored the reference group and it was hypothesized this could lead to differences in effect size estimates. The results from a full factorial ANOVA modeling relative bias indicated that the highest-level interaction with a moderate effect size was the interaction between multilevel DIF framework and which group the item favored.

In general, both the BMH95 and SIBTEST BSSE overestimated effect sizes for the focal group and underestimated effect sizes for the reference group. Likely, these results are in part due to the fact that there were numerous conditions which favored the reference group. Namely conditions when impact was introduced. The presence of impact could be confounded with larger magnitude DIF effects for the focal group.

When RMSE was modeled with the group an item favored, none of the interactions had significant and medium- to large-effect sizes. Only the main effects of social-unit cluster sample size and multilevel DIF framework were significant. In the case of RMSE, which group the item favored did not appear to have a significant and practical effect on the observed RMSE.

Implications. Significantly, there appears to be true differences in the performance of the three multilevel DIF frameworks under the various conditions adopted in this study. Of the two adjustment methods, the BMH95 appears to have outperformed the SIBTEST BSSE due to the higher observed power rates under the conditions of small magnitude DIF (0.3). While the multilevel Rasch model suffered from inflated Type I error rates under various conditions, different modeling decisions could have been made which may have lowered the Type I error and increased the power rates. Specifically, only four items were used as anchors while the other thirty-six were tested for DIF. In a simulation study conducted by Wen (2014), more confirmatory approaches where less items were tested for DIF outperformed the modeling approach which was adopted for this study.

However, the effect size results favor the multilevel Rasch model. This is not surprising given the nature of the three multilevel DIF frameworks which were compared. The adjustments made to the SIBTEST and Mantel-Haenszel do not adjust effect size estimates. The inclusion of an effect size comparison is instructive as it demonstrates the inferior performance of unilevel DIF approaches in multilevel data. It appears that the effect size results are not systemically biased rather generally imprecise.

The superior effect size estimates coupled with the flexibility of the multilevel Rasch model make it a likely model to be championed in future research. However, care should be taken when utilizing a three-level Rasch model to detect social-unit DIF in the presence of impact. The multilevel Rasch model over identified items as containing DIF and overestimated effect size estimates. It appears that in the presence of impact, DIF and impact were conflated. This was confirmed in the separate impact analysis, in which estimates of impact were found to be underestimated in the presence of impact.

With regards to the appropriateness of the studied multilevel DIF frameworks for multilevel validation, all methods are appropriate for use when the different levels of validation maintain separation. However, when interaction between levels is the case (e.g. when a true ecological model is used) only the multilevel Rasch model provides sufficient evidence. Therefore, researchers must choose the appropriate methodology based on statistical considerations but also on their theoretical understanding of multilevel test use and validation.

Assessing and finding social-unit level DIF will have implications for researchers. As highlighted by the theoretical results discussed in Chapter Four, main effect social-unit DIF can take two main forms. First, would be cases where information is gained but items are not considered bias. Studies of this sort are likely to add value for practitioners, researchers, and policy makers. On the other hand, are cases in which differences are due to construct-irrelevant variance and further judgments of bias are necessary. These analyses would likely be of interest to psychometricians and test developers and would be a crucial part of a multilevel validation process to ensure fairness.

The detection of social-unit DIF will have significant implications when aggregating scores. It is inadvisable to aggregate scores which are biased against a group. However, DIF at the item level does not equate to differential test functioning. Assuming items containing statistical DIF have been found to be biased, it will need to be determined how much item contamination and of what magnitude negatively impacts aggregate scores. Research of this nature will need to be undertaken in the future in order to offer guidelines to practitioners for the consideration of social-unit DIF.

Limitations

There are several limitations and future directions for this research. First, the method used to generate the data is overly simplistic. While multidimensional IRT was used to more realistically generate DIF in the data, the ability and nuisance dimension were not correlated. Therefore, results are more accurately attributed to the various factors related to DIF.

A significant limitation is the differences between anchor item selection across the three multilevel DIF frameworks. Selections were made in keeping with the literature on the various methods and for practicality. Numerous methods exist for building a purified subset and selecting anchor sets, most of which have been unexplored with these multilevel DIF frameworks. Testing multiple methods would have limited the other conditions of the simulation study and likely prohibited the comparative nature.

The selected anchor item and purified subset conditions make the comparability of the results questionable. The multilevel Rasch model may have been particularly handicapped by the use of only four anchor items, though this approach was in keeping with the current literature. While the SIBTEST BSSE method may have had overly strong results due to the large size of the anchor test. Analyses were conducted to confirm the extremely low Type I error and high power rates observed for the SIBTEST BSSE. It was found that when the anchor set decreased by as few as two items, the Type I error rate increased.

While the study of effect size measures under the three multilevel DIF frameworks was undertaken this work was limited. How effect sizes are measured varies across the three methods. The Beggs adjustment for the Mantel-Haenszel adjusts variance and therefore only affects the test for statistical significance. Similarly, the multilevel SIBTEST BSSE approach also only adjust the variance and therefore the test for statistical significance. However, since a bootstrapped $\hat{\beta}_{UNI}$ was calculated it was possible to compare the bootstrapped effect size measure, a minimal improvement over the initially calculated measure. Additionally, there is not a mathematical equation linking derived effect size estimates from the Mantel-Haenszel or Rasch model to SIBTEST. Rather an empirically derived relationship was used. However, the evaluation of effect size measures is an important step in understanding the practical significance of statistically significant DIF.

Lastly, this study only explores one type of multilevel DIF. When introducing the concept of multiple levels of data, it is possible for DIF to exist as main effects on both levels, interact across levels, or exist on the second level and vary across social-units. However, only the multilevel Rasch model is capable of teasing out these differing types of DIF. By researching the most simplistic version of social-unit DIF, this study is limited in its abilities to provide supporting evidence for an ecological model of validation as well as future frameworks which consider the interaction between nested levels.

Directions for Future Research

Future research should explicitly address some of the limitations outlined above. Regarding the overly simplistic data generation methods, simulation studies in general tend to suffer from this limitation (Luecht & Ackerman, 2018). Future studies may want to consider this work as a baseline condition and build upon it by studying the effects of varying the degree of correlation between the ability and nuisance dimensions.

Regarding the purified subtest and anchor item selection methodologies, future research should expand literature on anchor item selection in two-level models to three-level models. Current guidelines are adaptable for three-level research but need to concretely describe how to iteratively test for DIF in items at the second and third levels. Results of various methodologies will need to be presented so that a gold standard can be championed.

Regarding the comparability of results, future research should use the same selection criteria for the anchor items/purified subset to ensure comparability across methods. Likely this research will need to be undertaken after best practices have been devised for selecting anchor items in a three-level approach. Care should be taken to ensure the reasonableness of the anchor items/purified subset size. As research has demonstrated, larger anchor item sets yield better results (Wen, 2014). However, how realistic obtaining such large anchor item sets is should be proven.

Additionally, research should focus on expanding the conditions under which the different approaches are compared. Particularly, situations where impact is present as there has been minimal research on three-level Rasch models for DIF detection in impact and it proved to be a significant condition within this study.

The poorly estimated effect size estimates when using the BMH95 and SIBTEST BSSE highlight the need for modifications to be made to our unilevel DIF detection measures for multilevel data. Statistical flagging of DIF alone is not enough, practitioners require methods for accurately estimated the practical significance of DIF. The Mantel-Haenszel is a commonly used DIF detection method operationally and its multilevel adoption may be easier as many practitioners are familiar with it. Therefore, it would particularly benefit from effect size adjustments in the presence of multilevel data.

Future studies should examine the performance of the three multilevel DIF frameworks under different types of DIF. However, when considering the SIBTEST BSSE and BMH95 methods it may only be informative from the standpoint of understanding how severely impacted the frameworks are under these more complicated conditions.

This line of research is salient outside of bias studies, particularly as the desire to have finer grain information regarding item features, item difficulty, and the causes of bias grows. According the Launeanu and Hubley (2017) such inquires will allow the exploration of a culturally situated response process cycle.

As research on more complex forms of multilevel DIF is undertaken linking simulation studies to realistic operational scenarios is crucial for helping multilevel DIF studies gain traction operationally. These frameworks are more complex than their uni-level counterparts and offer significantly more information. While an argument can be made that our current education landscape necessitates such work, adoption of these frameworks will be swifter if they can be shown to be operationally beneficial. Thus, future research should investigate how multilevel DIF frameworks can add value for testing programs and practitioners. As an example, incorporating item level features could add efficiencies for item review processes. At the social-unit level, opportunity to learn variables related to funding and policy choices could be used in efficacy research. While not all the work is purely related to DIF analyses, bundling the endeavors will help to speed widespread adoption.

Lastly, future research should continue to link theoretical discussions of validation, accountability, and context with the use of multilevel DIF frameworks. There is a dearth of literature tackling both the theoretical and statistical sides of the multilevel validation discussion. Statistical studies in particular would be strengthened by providing more evidence of their own importance. Similarly, theoretical discussions should follow in the footsteps of Chen and Zumbo (2017) by providing example applied analyses of recommended methodologies.

Not only does applied multilevel DIF research need to be linked to theoretical work, it should be part of a growing body of research addressing use at the social-unit level and multilevel validation. When investigating more complex forms of DIF, researchers will have to address to what end. Is the interaction of item, person, and societal characteristics mandated by validation? If so, do the inferences we currently investigate and support adequately embody this research?

Given the nature of the ecological model for validation presented within this work it is unlikely that our current inferences will be adequate. In a scenario where there is interaction between levels of nesting there will likely need to be inferences explicitly for those interactions. At the very least, new studies and types of evidence will need to be documented to guide practitioners in their endeavors.

A fully interactional framework would represent a complex web that would likely be difficult to implement operationally. The ecological validation model and multilevel Interpretation/Use Argument fall short of such a lofty goal. At one end of the spectrum, the ecological validation model is interactional but it fails to provide a useful roadmap for practitioners. At the other end of the spectrum, the multilevel IUA is a practical approach to addressing validation operationally. However, while it addresses two levels of nesting it fails to account for their interaction.

Significantly more work must be done to visualize the multilevel validation process and to propose frameworks which are both theoretically sound and operationally useful. While advanced methodologies have made the work possible, we should not let our endeavors be driven solely by what we are capable of doing rather by what is necessitated to support test use and interpretation.

REFERENCES

Acar, T. (2012). Determination of a differential item functioning procedure using the hierarchical generalized linear model: A comparison study with logistic regression and likelihood ratio procedure. *Sage Open*, 1-8.

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*(1), 67-91.

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

Atar, B. (2007). *Differential item functioning analyses for mixed response data using IRT likelihood-ratio test, logistic regression, and GLLAMM procedures*. Unpublished dissertation: Florida State University.

Bachman, L. F., & Palmer, A.S. (2010). *Language assessment in practice*. London: Oxford University Press.

Begg, C. (1999). Analyzing k (2 3 2) tables under cluster sampling. *Biometrics*, *55*, 302-307.

Beimars, J. N., Walters, D. W., McClarty, K. L., & Miles, J. A. (2012). *Evidenced based standard setting: Establishing cut scores by integrating research evidence with expert content judgments*. Pearson Bulletin, January 2012, Issue 21. Retrieved October 10, 2017 from www.pearsonassessments.com.

Bennett, R. E., Kane, M., & Bridgeman, B. (2011). *Theory of action and validity argument in the context of through-course summative assessment.* Presented at the Invitational Research Symposium on Through-Course Summative Assessments. Atlanta, GA.

Binci, S. (2007). *Random effect differential item functional via hierarchical generalized linear model and generalized linear latent mixed model: A comparison of estimation methods.* Unpublished dissertation: Florida State University.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick (Eds). *Statistical Theories of Mental Test Scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Bolt, D.M. (2000). A SIBTEST approach to testing DIF hypotheses using experimentally designed test items. *Journal of Educational Measurement, 37,* 307-327.

Borsboom, D. & Wijsen, L. (2015). Frankenstein's validity monster: the value of keeping politics and science separated. *Assessment in Education: Principles, Policy & Practice, 23,* 281-283.

Brofenbrenner, U. (1994). Ecological models of human development. In t. Huston & T. N. Postlethwaith (Eds.), *International encyclopedia of education,* 2nd ed., Vol. 3 (pp. 1643-1647). New York, NY: Elsevier Science.

Burkes, L.L. (2009). *Identifying differential item functioning related to student socioeconomic status and investigating sources related to classroom opportunities to learn.* Unpublished dissertation: University of Delaware.

Cai, L. (2015). *Examining sources of gender DIF using cross-classification multilevel IRT models*. Unpublished thesis: University of Nebraska.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221-256). Westport, CT: Praeger.

Camilli, G. (1993). The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues?. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 397–417). Hillsdale, NJ: Erlbaum.

Camilli, G., & Penfield, D. A. (1997). Variance estimation for differential test functioning based on Mantel-Haenszel statistics. *Journal of Educational Measurement, 34,* 123-139.

Camilli, G., & Shepard, L. (1994). Methods for identifying biased test items. Newbury Park, CA: Sage.

Castellano, K.E. & Ho, A.D. (2013). *A Practitioner's Guide to Growth Models*. Council of Chief State School Officers.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing, 20,* 369-383.

Chalhoub-Deville, M. (2009). Standards-based assessment in the U.S.: Social and educational impact. In Taylor, L. and Weir, C. J. (Eds.), *Language Testing*

Matters: Investigating the wider social and educational impact of assessment (281-300). Cambridge: Cambridge University Press and Cambridge ESOL.

Chalhoub-Deville, M. (2016). Validity theory: Reform policies, accountability testing, and consequences. *Language Testing, 33*(4), 453-472.

Chapelle, C.A., Enright, M.K., Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3-13.

Chen, G., Mathie, J.E., & Bliese, P.D. (2004). A framework for conducting multilevel construct validation. In F.J. Yammarino & F. Dansereau (Eds.), *Research in multilevel issues: Multilevel issues in organizational behavior and processes* (pp. 273-303). Oxford, England: Elsevier.

Chen, J.H., Chen, C.T., & Shih, C.L. (2013). Improving the control of Type I error rate in assessing differential item functioning for hierarchical generalized linear model when impact is presented. *Applied Psychological Measurement,* 38(1), 18-36.

Chen, M.Y., Zumbo, B.D. (2017). Ecological framework of item responding as validity evidence: An application of multilevel DIF modeling using PISA data. In B.D. Zumbo & A.M. Hubley (Eds.), *Understanding and Investigating Response processes in Validation Research* (pp. 53-68). New York, NY: Springer.

Cheong, Y.F. (2006). Analysis of school context effects on differential item functioning using hierarchical generalized linear models. *International Journal of Testing*, 6(1), 57-79.

Cid, J. (2009). *Using explanatory item response models to examine the impact of linguistic features of a reading comprehension test on English language learners.* Unpublished dissertation: James Madison University.

Cizek, G.J.. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods, 17,* 31-43.

Cizek, G.J. (2016). Validating test score meaning and defending test score use: different aims, different methods. *Assessment in Education: Principles, Policy & Practice, 23*, 212-225.

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

Clauser, B., Mazor, K., & Hambleton, R. K. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel–Haenszel procedure. Applied Measurement in Education, 6, 269-279.

Clauser, B. E., Nungester, R. J., & Swaminathan, H. (1996). Improving the matching for DIF analysis by conditioning on both test scores and an educational background variable. *Journal of Educational Measurement*, 33, 453–464.

Coburn, C.E., Toure, J., & Yamashita, M. (2009). Evidence, interpretation, and persuasion: Instructional decision making at the district central office. *The Teachers College Record, 111,* 1115-1161.

Coburn, C.E. & Turner, E.O. (2012). The practice of data use: An introduction. *American Journal of Education, 118,* 99-111.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Second Edition. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cooper, H.M. & Hedges, L.V. (1994). *The handbook of research synthesis.* New York, NY: The Russell Sage Foundation.

Deville, C. & Chalhoub-Deville, M. (2011). Accountability-assessment under No Child Left Behind: Agenda, practice, and future. *Language Testing, 28,* 307-321.

Donoghue, J., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 137-166). Hillsdale, N J: Erlbaum.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Erlbaum.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23,* 355–368.

Embretson, S.E. (Ed.) (2010). *Measuring psychological constructs: Advances in model-based approaches*. Washington, D.C.: American Psychological Association Books.

ETS. (2014). *ETS standards for quality and fairness*. Retrieved on January 15, 2018 from https://www.ets.org/s/about/pdf/standards.pdf

Every Student Succeeds Act of 2015, Pub. L. No. 114-95, § 114 Stat. 1177 (2015-2016).

Finch, H. & French, B.F. (2010). Detecting differential item functioning of a course satisfaction instrument in the presence of multilevel data. *Journal of the First-Year Experience & Students in Transition, 22,* 27-47.

French, B.F. & Finch, W.H. (2015). Transforming SIBTEST to account for multilevel data structures. *Journal of Educational Measurement, 52*(2), 159-180.

French, B.F. & Finch, W.H. (2013). Extensions of Mantel-Haenszel for multilevel DIF detection. *Educational and Psychological Measurement*, *73*(4), 648-671.

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement,* 67, 373-393.

Gomez-Benito, J., Sireci, S., Padilla, J.-L., Hidalgo, D.M., Benitez, I. (2018). Differntial item functioning: Beyond validity evidence on internal structure. *Psicothema*, *30*(1), 1-15.

Guarino, C., Reckase, M., Stacy, B., and Wooldridge, J. (2015) Evaluating Specification Tests in the Context of Value-Added Models of Teacher Performance*, Journal of Research on Educational Effectiveness*, 8(1), 35-59.

Hambleton, R.K. & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. New York, NY: Springer Science & Business Media.

Haertel, E.H. (2013). *Reliability and validity of inferences about teachers based on student test scores*. William H. Angoff Memorial Lecture Series. Washington, D.C.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group randomized trials in education, Educational Evaluation and Policy Analysis, 29, 60-87.

Helms, C. (2018). More NC teachers are getting bigger bonuses in 2018. So why does controversy remain? *Charlotte Observer*. Retrieved from http://www.charlotteobserver.com/news/local/education/article197814844.html

Hidalgo, M.D. & Lopez-Pina, J.A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64*(6), 903-915.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In Wainer, H., & Braun, H. I. (Eds.), Test validity (pp. 129-145). Hillsdale, NJ: Erlbaum.

Hopmann, S.T. (2008). No child, no school, no state left behind: Schooling in the age of accountability. *Journal of Curriculum Studies*, *40*(4), 417-456.

Hubley, A.M. & Zumbo, B.D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research, 103,* 219-230.

Jak, S., Oort, F. J., & Dolan, C. V. (2013). A test for cluster bias: detecting violations of measurement invariance across clusters in multilevel data. *Structure Equation Modeling*, 20(2), 265-282.

Jodoin, M.G. & Gierl, M.J. (2001). Evaluating Type I error and power rates using an effect size measure with logistic regression procedure for DIF detection. *Applied Measurement in Education*, 14, 329-349.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*, 79–93.

Kamata, A., Chaimongkol, S., Genc, E., & Bilir, K. (April 2005). *Random-effect differential item functioning across group units by the hierarchical generalized linear model*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Kamata, A. & Cheong, Y.F. (2007). Multilevel Rasch models. In M. von Davier & C.H. Carstensen (Eds.), *Multivariate and Mixture Distribution Rasch Models: Extensions and Applications* (pp. 217-232). New York, NY: Springer.

Kane, M., Crooks, T.J., & Cohen, A.S. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5-17.

Kane, M. (2006). Validation. In R.Brennan (Ed.), Educational measurement (4th ed., pp. 17–64). Westport, CT: Greenwood Publishing.

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

Kim, S.H., Cohen, A.S., Alagoz, C., & Kim, S. (2007). DIF detection and effect size measures for polytomously scored items. *Journal of Educational Measurement, 44*(2), 93-116.

Kim, S.H., Cohen, A.S., & Park, T.H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement, 32*(3), 261-276.

Launeanu, M. & Hubley, A. M. (2017). A model building approach to examining response processes as a source of validity evidence for self-report items and measures. In B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and* *investigating response processes: Advances in validation research* (pp. 115-136). New York, NY: Springer.

Levinson, A. & Schauss, A. (2017). *Class size: Allotments, requirements, and reporting per HB13 (S.L. 2017-9).* Retrieved on May 14, 2018 from http://www.ncpublicschools.org/docs/fbs/accounting/hb13classsize.pdf

Li, H.H., & Stout, W.F. (1996). A new procedure for detection of crossing DIF. *Psychometrika, 61*, 647-677.

Li, Qin, Lei. (2017). An examination of the instructional sensitivity of the TIMSS math items: A hierarchical differential item functioning approach. *Educational Assessment, 22*, 1-17.

Linn, R. (2005). *Test-based Educational Accountability in the Era of No Child Left Behind*. The Regents of the University of California: CSE Report 651.

Lissitz, R.W. & Samuelsen. K. (2009). Dialogue on Validity: A Suggested Change in Terminology and Emphasis Regarding Validity and Education. *Educational Researcher, 36,* 437-448.

Liu, Q. (2011). *Item purification in differential item functioning using generalized linear mixed models*. Unpublished dissertation: Florida State University.

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

Luecht, R. & Ackerman, T.A. (2018). A technical note on IRT simulation studies: Dealing with truth, estimates, observed data, and residuals. *Educational Measurement: Issues and Practice, 00,* 1-12.

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

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.

Mathis, W.J. & Truijillo, T.M. (2016). *Lessons from NCLB for the Every Student Succeeds Act*. Boulder, CO: National Education Policy Center. Retrieved December 14, 2017 from http://nepc.colorado.edu/publication/lessons-from-NCLB

Mazzeo, C. (2001). Frameworks of state: Assessment policy in historical perspective. *Teachers College Record, 103*, 367-397.

McClarty, K. L., Way, W. D., Porter, A. C., Beimers, J. N., & Miles, J. A. (2013). Evidence-Based standard setting: Establishing a validity framework for cut scores. *Educational Researcher, 42*, 78-88.

Mellenbergh, G. J. (1982). Contingency table methods for assessing item bias. *Journal of Educational Statistics, 7,* 105-118.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*(4), 297-334.

Monahan, P. O., McHorney, C. A., Stump, T. E., & Perkins, A. J. (2007). Odds ratio, Delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics*, *32,* 92–109.

Moss, P. (2007). Reconstructing validity. *Educational Researcher, 36,* 470-476.

Moss, P. (2016). Shifting the focus of validity for test use. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 236-251.

Moulton, B. (1986). Random group effects and the precision of regression estimates. *Journal of Econometrics*, *32*, 385–397.

National Center for Education Statistics. (2012). *Schools and staffing survey*. Retrieved on May 14, 2018 from https://nces.ed.gov/surveys/sass/tables/sass1112_2013314_t1s_007.asp

NCDPI. (2013). *End-of-grade assessments: Number of items, item types, and estimated test time (minutes).* Retrieved on May 15, 2018 from http://www.ncpublicschools.org/docs/accountability/testing/eog/eogadmininfo13.pdf.

NCDPI. (2017a). *Accountability brief*. Retrieved on May 1, 2018 from http://www.ncpublicschools.org/docs/accountability/reporting/acctbrf17.pdf.

NCDPI. (2017b). *Employee salary and benefits manual: Section C school-based administrators, principals, and assistant principals*. Retrieved on May 1, 2018 from http://www.ncpublicschools.org/docs/fbs/finance/salary/ salarysectionscandd.pdf.

NCDPI. (2017c). *Fiscal year 2017-2018 North Carolina public school salary schedules.* Retrieved on May 1, 2018 from http://www.ncpublicschools.org/docs/fbs/finance/ salary/schedules/2017-18schedules.pdf.

NCDPI. (2017d). *Performance and growth of North Carolina public schools: Executive summary of statistical results.* Retrieved on May 1, 2018 from http://www.ncpublicschools.org/docs/accountability/reporting/2017/ documentation/exsumm17.pdf

No Child Left Behind Act of 2001, Pub. L. No. 107–110, 115 Stat. 1425 (2002).

O'Malley, K., Keng, L., & Miles, J. (2012). From Z to A: Using validity evidence to set performance standards. In G. J. Cizek (Ed.), *Setting Performance Standards: Foundation, Methods and Innovations* (pp. 237-260). New York, NY: Routledge.

Penfield, R.D., & Camilli, G. (2007). Differential item functioning and item bias. In S. Sinharay & C.R. Rao (Eds.), *Handbook of Statistics, Volume 26: Psychometrics* (pp. 125-167). New York, NY: Elsevier.

Potenza, M.T. & Dorans, N.J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19,* 23-37.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53,* 495-502.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Statistics for Social and Behavioral Sciences.

Roussos, L.A., Schnipke, D.L., & Pashley, P.J. (1999). A generalized formula for the Mantel-Haenszel differential item functioning parameter. *Journal of Educational and Behavioral Statistics, 24*(3), 293-322.

Roussos, L.A. & Stout, w. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20,* 355-371.

Sahlberg, P. (2011). *Global educational reform movement is here!* Retrieved January 15, 2018 from http://pasisahlberg.com/global-educational-reform-movement-is-here/

Shealy, R. T., & Stout, W. F. (1993). An item response theory model for test bias and differential test functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197-239). Hillsdale, NJ: Erlbaum.

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*, 184-199.

Sireci, S.G. (2016). On the validity of useless tests. *Assessment in Education: Principles, Policy & Practice, 23,* 226-235,

State Board of Education. (2017). Report to the North Carolina General Assembly: State funded personnel/merit – based bonuses to local education agency. *SL 2016-94 Section 36.1A(c).* Retrieved on May 1, 2018 from https://www.google.com/search?q=Report+to+the+North+Carolina+General+Assembly%3A+State+funded+personnel%2Fmerit+%E2%80%93+based+bonuses+to+local+education+agency&oq=Report+to+the+North+Carolina+General+Assembly%3A+State+funded+personnel%2Fmerit+%E2%80%93+based+bonuses+to+local+education+agency&aqs=chrome..69i57.375j0j7&sourceid=chrome&ie=UTF-8.

State Board of Education. (2015). *North Carolina teacher evaluation process.* Retrieved on May 5, 2018 from http://www.ncpublicschools.org/docs/effectiveness-model/ncees/instruments/teach-eval-manual.pdf.

Smith, W.C. (2017). National testing policies and educator based testing for accountability. *OECD Journal*: *Economic Studies*, February 2017.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In Wainer, H., & Braun, H. (Eds.), Test validity (pp. 147-169). Hillsdale, NJ: Erlbaum.

Toulmin, S. (1958). *The uses of argument.* Cambridge: Cambridge University Press.

Volante, L. (2007). *Evaluating test-based accountability perspectives: An international perspective.* Paper presented at the Association for Educational Assessment – Europe, Stockholm, Sweden.

Walker, C.M. (2011). What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. *Journal of Psychoeducational Assessment*, *29*, 364-376.

Wang, W.-C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *Journal of Experimental Education*, *72*, 221-261.

Wang, W. C., Shih, C. L., & Sun, G. W. (2012). The DIF-free-then-DIF strategy for the assessment of differential item functioning. *Educational and Psychological Measurement, 72,* 687-708.

Wang, W.-C., & Su, Y.-H. (2004). Effects of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel-Haenszel method. *Applied Measurement in Education, 17,* 113-144.

Wang, W. C., & Yeh, Y. L. (2003). Effects of anchor item methods on differential item functioning detection with the Likelihood Ratio Test. *Applied Psychological Measurement, 27,* 479-498.

Wen, Y. (2014). *DIF analyses in multilevel data: Identification and effects on ability estimates.* Unpublished dissertation, University of Wisconsin-Milwaukee.

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

Zhang, J. & Boos, D.D. (1997). Mantel-Haenszel test statistics for correlated binary data. *Biometrics*, 1185-1198.

Zhu, X.S., Rupp, A.A., & Gao, J. (2011). Differential item functioning analyses in large scale educational surveys: Key concepts and modeling approaches for secondary analysts. *Journal of Research in Education Sciences, 56,* 91-127.

Zieky, M. (1993). DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Hillsdale, NJ: Erlbaum.

Zumbo, B.D. (2009). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4,* 223-233.

Zumbo, B.D. & Forer, B. (2011). Testing and measurement from a multilevel view: Psychometrics and validation. In J.A. Bovaird, K.F. Geisinger, & C.W. Buckendahl (Eds.), *High Stakes Testing in Education: Science and Practice in K-12 Settings* (pp. 177-190). Washington, D.C.: American Psychological Association.

Zumbo, B. D., & Hubley, A. M. (2003). Item bias. In Rocío Fernández-Ballesteros (Ed.). *Encyclopedia of Psychological Assessment* (p. 505-509). Sage Press, Thousand Oaks, CA.

Zumbo, B.D., & Hubley, A.M. (2016). Bringing consequences and side effects of testing and assessment to the foreground. *Assessment in Education: Principles, Policy & Practice, 23,* 299-303.

Zumbo, B.D., Liu, Y., Wu, A.D., Shear, B.R., Olvera, O.L., Ark, A., & Ark, T.K. (2015). A methodology for Zumbo's third generation DIF analyses and the ecology of item responding. *Language Assessment Quarterly, 12,* 136-151.

Zwick, R., (2012, April). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement.* Educational Testing Services Research Report RR-12-08. Retrieved on December 10, 2018 from https://www.ets.org/Media/Research/pdf/RR-12-08.pdf

APPENDIX A

TYPE I ERROR AND POWER RESULTS

Table A.1

The ANOVA of Type I Error Rates

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Magnitude DIF	3	0.000	0.000	1.286	0.299	0.001
Number of Clusters	1	0.003	0.003	173.850	0.000	0.003
R:F Ratio	1	0.000	0.000	12.442	0.002	0.001
Impact Presence	1	0.006	0.006	342.078	0.000	**0.053**
ICC Magnitude	1	0.000	0.000	4.847	0.036	0.000
Multilevel DIF Framework	2	0.056	0.028	1571.172	0.000	**0.298**
DIF Magnitude x Number of Clusters	3	0.000	0.000	1.629	0.206	0.001
DIF Magnitude x R:F Ratio	3	0.000	0.000	2.354	0.094	0.001
DIF Magnitude x Impact Presence	3	0.000	0.000	7.648	0.001	0.004
DIF Magnitude x ICC Magnitude	3	0.004	0.001	78.529	0.000	0.031
DIF Magnitude x Multilevel DIF Framework	6	0.000	0.000	3.019	0.022	0.003
Number of Clusters x R:F Ratio	1	0.000	0.000	1.534	0.226	0.000
Number of Clusters x Impact Presence	1	0.002	0.002	100.065	0.000	0.014
Number of Clusters x ICC Magnitude	1	0.000	0.000	2.040	0.165	0.000
Number of Clusters x Multilevel DIF Framework	2	0.033	0.016	912.032	0.000	**0.117**
R:F Ratio x Impact Presence	1	0.000	0.000	0.481	0.494	0.000
R:F Ratio x ICC Magnitude	1	0.000	0.000	6.685	0.015	0.000
R:F Ratio x Multilevel DIF Framework	2	0.000	0.000	0.929	0.407	0.000
Impact Presence x ICC Magnitude	1	0.000	0.000	0.117	0.735	0.000
Impact Presence x Multilevel DIF Framework	2	0.010	0.005	274.183	0.000	**0.062**
ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	2.035	0.150	0.005

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
DIF Magnitude x Number of Clusters x R:F Ratio	3	0.000	0.000	1.007	0.405	0.000
DIF Magnitude x Number of Clusters x Impact Presence	3	0.000	0.000	3.074	0.045	0.001
DIF Magnitude x Number of Clusters x ICC Magnitude	3	0.002	0.001	41.869	0.000	0.011
DIF Magnitude x Number of Clusters x Multilevel DIF Framework	6	0.000	0.000	0.793	0.584	0.001
DIF Magnitude x R:F Ratio x Impact Presence	3	0.000	0.000	6.126	0.003	0.003
DIF Magnitude x R:F Ratio x ICC Magnitude	3	0.000	0.000	6.126	0.003	0.002
DIF Magnitude x R:F Ratio x Multilevel DIF Framework	6	0.000	0.000	0.469	0.825	0.000
DIF Magnitude x Impact Presence x ICC Magnitude	3	0.003	0.001	64.725	0.000	0.027
DIF Magnitude x Impact Presence x Multilevel DIF Framework	6	0.000	0.000	1.186	0.343	0.001
DIF Magnitude x ICC Magnitude x Multilevel DIF Framework	6	0.005	0.001	47.668	0.000	0.030
Number of Clusters x R:F Ratio x Impact Presence	1	0.000	0.000	3.264	0.082	0.000
Number of Clusters x R:F Ratio x ICC Magnitude	1	0.000	0.000	6.898	0.014	0.001
Number of Clusters x R:F Ratio x Multilevel DIF Framework	2	0.000	0.000	0.954	0.398	0.000
Number of Clusters x Impact Presence x ICC Magnitude	1	0.000	0.000	2.233	0.147	0.001
Number of Clusters x Impact Presence x Multilevel DIF Framework	2	0.003	0.001	79.384	0.000	0.018
Number of Clusters x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.527	0.596	0.000
R:F Ratio x Impact Presence x ICC Magnitude	1	0.000	0.000	4.879	0.036	0.000
R:F Ratio x Impact Presence x Multilevel DIF Framework	2	0.000	0.000	0.970	0.392	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
R:F Ratio x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	2.308	0.119	0.000
Impact Presence x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	4.964	0.015	0.001
DIF Magnitude x Number of Clusters x R:F Ratio x Impact Presence	3	0.000	0.000	3.811	0.021	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x ICC Magnitude	3	0.000	0.000	0.467	0.708	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Multilevel DIF Framework	6	0.000	0.000	1.468	0.226	0.000
DIF Magnitude x Number of Clusters x Impact Presence x ICC Magnitude	3	0.001	0.000	22.734	0.000	0.008
DIF Magnitude x Number of Clusters x Impact Presence x Multilevel DIF Framework	6	0.000	0.000	1.692	0.161	0.000
DIF Magnitude x Number of Clusters x ICC Magnitude x Multilevel DIF Framework	4	0.001	0.000	20.472	0.000	0.010
DIF Magnitude x R:F Ratio x Impact Presence x ICC Magnitude	3	0.000	0.000	7.328	0.001	0.002
DIF Magnitude x R:F Ratio x Impact Presence x Multilevel DIF Framework	6	0.000	0.000	2.109	0.085	0.001
DIF Magnitude x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	3	0.000	0.000	3.078	0.044	0.001
DIF Magnitude x Impact Presence x ICC Magnitude x Multilevel DIF Framework	3	0.005	0.002	88.152	0.000	0.033
Number of Clusters x R:F Ratio x Impact Presence x ICC Magnitude	1	0.000	0.000	0.108	0.745	0.000
Number of Clusters x R:F Ratio x Impact Presence x Multilevel DIF Framework	2	0.000	0.000	0.575	0.569	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Number of Clusters x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	2.184	0.132	0.001
Number of Clusters x Impact Presence x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	3.481	0.045	0.001
R:F Ratio x Impact Presence x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.245	0.785	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Impact Presence x ICC Magnitude	3	0.000	0.000	0.290	0.833	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Impact Presence x Multilevel DIF Framework	3	0.000	0.000	0.593	0.625	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	3	0.000	0.000	0.187	0.904	0.000
DIF Magnitude x Number of Clusters x Impact Presence x ICC Magnitude x Multilevel DIF Framework	3	0.001	0.000	21.612	0.000	0.008
DIF Magnitude x R:F Ratio x Impact Presence x ICC Magnitude x Multilevel DIF Framework	3	0.000	0.000	2.488	0.082	0.001
Number of Clusters x R:F Ratio x Impact Presence x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.175	0.840	0.000
Error	27	0.000	0.000			
Total	192	0.142				

Table A.2

The ANOVA of Power Rates

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Magnitude DIF	2	1.958	0.979	64232.570	0.000	**0.559**
Number of Clusters	1	0.373	0.373	24450.660	0.000	**0.106**
R:F Ratio	1	0.092	0.092	6065.470	0.000	0.026
Impact Presence	1	0.002	0.002	135.353	0.000	0.001
ICC Magnitude	1	0.002	0.002	113.166	0.000	0.000
Multilevel DIF Framework	2	0.082	0.041	2679.235	0.000	0.023
DIF Magnitude x Number of Clusters	2	0.445	0.223	14611.250	0.000	**0.127**
DIF Magnitude x R:F Ratio	2	0.103	0.051	3375.738	0.000	0.029
DIF Magnitude x Impact Presence	2	0.003	0.001	83.624	0.001	0.001
DIF Magnitude x ICC Magnitude	2	0.004	0.002	138.313	0.000	0.001
DIF Magnitude x Multilevel DIF Framework	4	0.141	0.035	2317.257	0.000	0.040
Number of Clusters x R:F Ratio	1	0.007	0.007	448.133	0.000	0.002
Number of Clusters x Impact Presence	1	0.000	0.000	4.710	0.096	0.000
Number of Clusters x ICC Magnitude	1	0.000	0.000	2.796	0.170	0.000
Number of Clusters x Multilevel DIF Framework	2	0.069	0.035	2276.765	0.000	0.020
R:F Ratio x Impact Presence	1	0.000	0.000	0.032	0.867	0.000
R:F Ratio x ICC Magnitude	1	0.000	0.000	0.285	0.622	0.000
R:F Ratio x Multilevel DIF Framework	2	0.018	0.009	606.294	0.000	0.005
Impact Presence x ICC Magnitude	1	0.000	0.000	0.011	0.920	0.000
Impact Presence x Multilevel DIF Framework	2	0.000	0.000	13.914	0.016	0.000
ICC Magnitude x Multilevel DIF Framework	2	0.002	0.001	79.372	0.001	0.001
DIF Magnitude x Number of Clusters x R:F Ratio	2	0.003	0.002	100.786	0.000	0.001
DIF Magnitude x Number of Clusters x Impact Presence	2	0.000	0.000	1.771	0.281	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
DIF Magnitude x Number of Clusters x ICC Magnitude	2	0.000	0.000	0.085	0.920	0.000
DIF Magnitude x Number of Clusters x Multilevel DIF Framework	4	0.141	0.035	2319.338	0.000	0.040
DIF Magnitude x R:F Ratio x Impact Presence	2	0.000	0.000	3.571	0.129	0.000
DIF Magnitude x R:F Ratio x ICC Magnitude	2	0.000	0.000	2.077	0.241	0.000
DIF Magnitude x R:F Ratio x Multilevel DIF Framework	4	0.023	0.006	376.217	0.000	0.007
DIF Magnitude x Impact Presence x ICC Magnitude	2	0.000	0.000	2.692	0.182	0.000
DIF Magnitude x Impact Presence x Multilevel DIF Framework	4	0.001	0.000	22.804	0.005	0.000
DIF Magnitude x ICC Magnitude x Multilevel DIF Framework	4	0.002	0.001	33.613	0.002	0.001
Number of Clusters x R:F Ratio x Impact Presence	1	0.000	0.000	2.796	0.170	0.000
Number of Clusters x R:F Ratio x ICC Magnitude	1	0.000	0.000	8.720	0.042	0.000
Number of Clusters x R:F Ratio x Multilevel DIF Framework	2	0.006	0.003	192.020	0.000	0.002
Number of Clusters x Impact Presence x ICC Magnitude	1	0.000	0.000	5.682	0.076	0.000
Number of Clusters x Impact Presence x Multilevel DIF Framework	2	0.000	0.000	2.648	0.185	0.000
Number of Clusters x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	11.610	0.022	0.000
R:F Ratio x Impact Presence x ICC Magnitude	1	0.000	0.000	0.923	0.391	0.000
R:F Ratio x Impact Presence x Multilevel DIF Framework	2	0.000	0.000	1.554	0.317	0.000
R:F Ratio x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	10.223	0.027	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Impact Presence x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	11.256	0.023	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Impact Presence	2	0.000	0.000	7.429	0.045	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x ICC Magnitude	2	0.000	0.000	2.401	0.206	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Multilevel DIF Framework	4	0.016	0.004	270.510	0.000	0.005
DIF Magnitude x Number of Clusters x Impact Presence x ICC Magnitude	2	0.000	0.000	7.103	0.048	0.000
DIF Magnitude x Number of Clusters x Impact Presence x Multilevel DIF Framework	4	0.000	0.000	8.082	0.034	0.000
DIF Magnitude x Number of Clusters x ICC Magnitude x Multilevel DIF Framework	4	0.001	0.000	10.403	0.022	0.000
DIF Magnitude x R:F Ratio x Impact Presence x ICC Magnitude	2	0.000	0.000	6.771	0.052	0.000
DIF Magnitude x R:F Ratio x Impact Presence x Multilevel DIF Framework	4	0.000	0.000	2.097	0.245	0.000
DIF Magnitude x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	4	0.000	0.000	3.448	0.129	0.000
DIF Magnitude x Impact Presence x ICC Magnitude x Multilevel DIF Framework	4	0.000	0.000	2.681	0.181	0.000
Number of Clusters x R:F Ratio x Impact Presence x ICC Magnitude	1	0.000	0.000	21.723	0.010	0.000
Number of Clusters x R:F Ratio x Impact Presence x Multilevel DIF Framework	2	0.000	0.000	2.982	0.161	0.000
Number of Clusters x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	3.438	0.135	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Number of Clusters x Impact Presence x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.248	0.791	0.000
R:F Ratio x Impact Presence x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	3.934	0.114	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Impact Presence x ICC Magnitude	2	0.001	0.001	45.328	0.002	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Impact Presence x Multilevel DIF Framework	4	0.000	0.000	1.908	0.273	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	4	0.000	0.000	3.453	0.129	0.000
DIF Magnitude x Number of Clusters x Impact Presence x ICC Magnitude x Multilevel DIF Framework	4	0.000	0.000	4.471	0.088	0.000
DIF Magnitude x R:F Ratio x Impact Presence x ICC Magnitude x Multilevel DIF Framework	4	0.000	0.000	5.089	0.072	0.000
Number of Clusters x R:F Ratio x Impact Presence x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	2.572	0.191	0.000
Error	4	0.000	0.000			
Total	143	3.502				

Table A.3

The ANOVA Results for Power Including Favored Group

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Favored Group	1	74.83	74.83	130.70	0.00	0.00
Magnitude DIF	2	5534.37	2767.18	4832.87	0.00	0.01
Number of Clusters	1	2544.10	2544.10	4443.26	0.00	0.00
R:F Ratio	1	522.58	522.58	912.68	0.00	0.00
Impact Presence	1	6.03	6.03	10.52	0.03	0.00
ICC Magnitude	1	6.52	6.52	11.39	0.03	0.00
Multilevel DIF Framework	2	502929.30	251464.70	439181.80	0.00	**0.91**
Favored Group x Magnitude of DIF	2	71.74	35.87	62.65	0.00	0.00
Favored Group x Number of Clusters	1	14.96	14.96	26.13	0.01	0.00
Favored Group x R:F Ratio	1	2.13	2.13	3.73	0.13	0.00
Favored Group x Impact Presence	1	58.06	58.06	101.40	0.00	0.00
Favored Group x ICC Magnitude	1	0.26	0.26	0.46	0.53	0.00
Favored Group x Multilevel DIF Framework	2	150.26	75.13	131.22	0.00	0.00
DIF Magnitude x Number of Clusters	2	3664.88	1832.44	3200.35	0.00	0.01
DIF Magnitude x R:F Ratio	2	791.27	395.64	690.98	0.00	0.00
DIF Magnitude x Impact Presence	2	5.50	2.75	4.80	0.09	0.00
DIF Magnitude x ICC Magnitude	2	3.38	1.69	2.95	0.16	0.00
DIF Magnitude x Multilevel DIF Framework	4	10441.26	2610.32	4558.90	0.00	0.02
Number of Clusters x R:F Ratio	1	342.33	342.33	597.88	0.00	0.00
Number of Clusters x Impact Presence	1	0.03	0.03	0.05	0.84	0.00
Number of Clusters x ICC Magnitude	1	0.05	0.05	0.09	0.78	0.00
Number of Clusters x Multilevel DIF Framework	2	4941.14	2470.57	4314.84	0.00	0.01

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
R:F Ratio x Impact Presence	1	0.00	0.00	0.01	0.94	0.00
R:F Ratio x ICC Magnitude	1	0.15	0.15	0.26	0.64	0.00
R:F Ratio x Multilevel DIF Framework	2	1000.31	500.15	873.52	0.00	0.00
Impact Presence x ICC Magnitude	1	171.32	171.32	299.21	0.00	0.00
Impact Presence x Multilevel DIF Framework	2	11.49	5.74	10.03	0.03	0.00
ICC Magnitude x Multilevel DIF Framework	2	12.15	6.08	10.61	0.03	0.00
Favored Group x DIF Magnitude x Number of Clusters	2	7.24	3.62	6.32	0.06	0.00
Favored Group x DIF Magnitude x R:F Ratio	2	17.56	8.78	15.33	0.01	0.00
Favored Group x DIF Magnitude x Impact Presence	2	71.29	35.64	62.25	0.00	0.00
Favored Group x DIF Magnitude x ICC Magnitude	2	0.71	0.35	0.62	0.58	0.00
Favored Group x DIF Magnitude x Multilevel DIF Framework	4	144.35	36.09	63.02	0.00	0.00
Favored Group x Number of Clusters x R:F Ratio	1	9.22	9.22	16.10	0.02	0.00
Favored Group x Number of Clusters x Impact Presence	1	8.05	8.05	14.05	0.02	0.00
Favored Group x Number of Clusters x ICC Magnitude	1	1.34	1.34	2.34	0.20	0.00
Favored Group x Number of Clusters x Multilevel DIF Framework	2	29.96	14.98	26.16	0.01	0.00
Favored Group x R:F Ratio x Impact Presence	1	0.20	0.20	0.35	0.59	0.00
Favored Group x R:F Ratio x ICC Magnitude	1	0.21	0.21	0.37	0.57	0.00

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Favored Group x R:F Ratio x Multilevel DIF Framework	2	4.38	2.19	3.82	0.12	0.00
Favored Group x Impact Presence x ICC Magnitude	1	0.11	0.11	0.20	0.68	0.00
Favored Group x Impact Presence x Multilevel DIF Framework	2	117.09	58.54	102.25	0.00	0.00
Favored Group x ICC Magnitude x Multilevel DIF Framework	2	0.46	0.23	0.40	0.69	0.00
DIF Magnitude x Number of Clusters x R:F Ratio	2	484.56	242.28	423.14	0.00	0.00
DIF Magnitude x Number of Clusters x Impact Presence	2	2.81	1.40	2.45	0.20	0.00
DIF Magnitude x Number of Clusters x ICC Magnitude	2	3.74	1.87	3.27	0.14	0.00
DIF Magnitude x Number of Clusters x Multilevel DIF Framework	4	7142.77	1785.69	3118.70	0.00	0.01
DIF Magnitude x R:F Ratio x Impact Presence	2	0.76	0.38	0.67	0.56	0.00
DIF Magnitude x R:F Ratio x ICC Magnitude	2	0.14	0.07	0.12	0.89	0.00
DIF Magnitude x R:F Ratio x Multilevel DIF Framework	4	1523.96	380.99	665.40	0.00	0.00
DIF Magnitude x Impact Presence x ICC Magnitude	2	362.20	181.10	316.29	0.00	0.00
DIF Magnitude x Impact Presence x Multilevel DIF Framework	4	10.37	2.59	4.53	0.09	0.00
DIF Magnitude x ICC Magnitude x Multilevel DIF Framework	4	6.14	1.54	2.68	0.18	0.00
Number of Clusters x R:F Ratio x Impact Presence	1	0.07	0.07	0.12	0.75	0.00
Number of Clusters x R:F Ratio x ICC Magnitude	1	0.32	0.32	0.56	0.50	0.00

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Number of Clusters x R:F Ratio x Multilevel DIF Framework	2	678.19	339.10	592.23	0.00	0.00
Number of Clusters x Impact Presence x ICC Magnitude	1	153.83	153.83	268.66	0.00	0.00
Number of Clusters x Impact Presence x Multilevel DIF Framework	2	0.07	0.03	0.06	0.94	0.00
Number of Clusters x ICC Magnitude x Multilevel DIF Framework	2	0.09	0.04	0.08	0.93	0.00
R:F Ratio x Impact Presence x ICC Magnitude	1	190.61	190.61	332.90	0.00	0.00
R:F Ratio x Impact Presence x Multilevel DIF Framework	2	0.01	0.00	0.01	0.99	0.00
R:F Ratio x ICC Magnitude x Multilevel DIF Framework	2	0.27	0.13	0.23	0.80	0.00
Impact Presence x ICC Magnitude x Multilevel DIF Framework	2	343.60	171.80	300.05	0.00	0.00
Favored Group x DIF Magnitude x Number of Clusters x R:F Ratio	2	40.63	20.31	35.48	0.00	0.00
Favored Group x DIF Magnitude x Number of Clusters x Impact Presence	2	3.89	1.95	3.40	0.14	0.00
Favored Group x DIF Magnitude x Number of Clusters x ICC Magnitude	2	8.30	4.15	7.25	0.05	0.00
Favored Group x DIF Magnitude x Number of Clusters x Multilevel DIF Framework	4	14.55	3.64	6.35	0.05	0.00
Favored Group x DIF Magnitude x R:F Ratio x Impact Presence	2	4.25	2.12	3.71	0.12	0.00
Favored Group x DIF Magnitude x R:F Ratio x ICC Magnitude	2	2.68	1.34	2.34	0.21	0.00

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Favored Group x DIF Magnitude x R:F Ratio x Multilevel DIF Framework	4	35.53	8.88	15.51	0.01	0.00
Favored Group x DIF Magnitude x Impact Presence x ICC Magnitude	2	0.02	0.01	0.02	0.98	0.00
Favored Group x DIF Magnitude x Impact Presence x Multilevel DIF Framework	4	143.84	35.96	62.80	0.00	0.00
Favored Group x DIF Magnitude x ICC Magnitude x Multilevel DIF Framework	4	1.53	0.38	0.67	0.65	0.00
Favored Group x Number of Clusters x R:F Ratio x Impact Presence	1	5.43	5.43	9.48	0.04	0.00
Favored Group x Number of Clusters x R:F Ratio x ICC Magnitude	1	1.18	1.18	2.05	0.23	0.00
Favored Group x Number of Clusters x R:F Ratio x Multilevel DIF Framework	2	18.59	9.30	16.23	0.01	0.00
Favored Group x Number of Clusters x Impact Presence x ICC Magnitude	1	3.81	3.81	6.65	0.06	0.00
Favored Group x Number of Clusters x Impact Presence x Multilevel DIF Framework	2	16.29	8.14	14.22	0.02	0.00
Favored Group x Number of Clusters x ICC Magnitude x Multilevel DIF Framework	2	2.69	1.35	2.35	0.21	0.00
Favored Group x R:F Ratio x Impact Presence x ICC Magnitude	1	0.22	0.22	0.38	0.57	0.00
Favored Group x R:F Ratio x Impact Presence x Multilevel DIF Framework	2	0.41	0.21	0.36	0.72	0.00

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Favored Group x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	2	0.40	0.20	0.35	0.73	0.00
Favored Group x Impact Presence x ICC Magnitude x Multilevel DIF Framework	2	0.22	0.11	0.19	0.83	0.00
DIF Magnitude x Number of Clusters x R:F Ratio x Impact Presence	2	1.64	0.82	1.43	0.34	0.00
DIF Magnitude x Number of Clusters x R:F Ratio x ICC Magnitude	2	2.54	1.27	2.22	0.22	0.00
DIF Magnitude x Number of Clusters x R:F Ratio x Multilevel DIF Framework	4	970.56	242.64	423.77	0.00	0.00
DIF Magnitude x Number of Clusters x Impact Presence x ICC Magnitude	2	325.87	162.93	284.56	0.00	0.00
DIF Magnitude x Number of Clusters x Impact Presence x Multilevel DIF Framework	4	5.72	1.43	2.50	0.20	0.00
DIF Magnitude x Number of Clusters x ICC Magnitude x Multilevel DIF Framework	4	7.38	1.84	3.22	0.14	0.00
DIF Magnitude x R:F Ratio x Impact Presence x ICC Magnitude	2	334.67	167.34	292.25	0.00	0.00
DIF Magnitude x R:F Ratio x Impact Presence x Multilevel DIF Framework	4	1.49	0.37	0.65	0.66	0.00
DIF Magnitude x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	4	0.26	0.06	0.11	0.97	0.00
DIF Magnitude x Impact Presence x ICC Magnitude x Multilevel DIF Framework	4	724.03	181.01	316.13	0.00	0.00

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Number of Clusters x R:F Ratio x Impact Presence x ICC Magnitude	1	200.18	200.18	349.61	0.00	0.00
Number of Clusters x R:F Ratio x Impact Presence x Multilevel DIF Framework	2	0.13	0.06	0.11	0.90	0.00
Number of Clusters x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	2	0.65	0.33	0.57	0.61	0.00
Number of Clusters x Impact Presence x ICC Magnitude x Multilevel DIF Framework	2	307.01	153.50	268.09	0.00	0.00
R:F Ratio x Impact Presence x ICC Magnitude x Multilevel DIF Framework	2	381.39	190.69	333.05	0.00	0.00
Favored Group x DIF Magnitude x Number of Clusters x R:F Ratio x Impact Presence	2	21.81	10.90	19.04	0.01	0.00
Favored Group x DIF Magnitude x Number of Clusters x R:F Ratio x ICC Magnitude	2	6.16	3.08	5.38	0.07	0.00
Favored Group x DIF Magnitude x Number of Clusters x R:F Ratio x Multilevel DIF Framework	4	81.72	20.43	35.68	0.00	0.00
Favored Group x DIF Magnitude x Number of Clusters x Impact Presence x ICC Magnitude	2	9.54	4.77	8.33	0.04	0.00
Favored Group x DIF Magnitude x Number of Clusters x Impact Presence x Multilevel DIF Framework	4	7.90	1.97	3.45	0.13	0.00
Favored Group x DIF Magnitude x Number of Clusters x ICC Magnitude	4	16.87	4.22	7.37	0.04	0.00

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
x Multilevel DIF Framework						
Favored Group x DIF Magnitude x R:F Ratio x Impact Presence x ICC Magnitude	2	0.10	0.05	0.09	0.92	0.00
Favored Group x DIF Magnitude x R:F Ratio x Impact Presence x Multilevel DIF Framework	4	8.60	2.15	3.75	0.11	0.00
Favored Group x DIF Magnitude x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	4	5.27	1.32	2.30	0.22	0.00
Favored Group x DIF Magnitude x Impact Presence x ICC Magnitude x Multilevel DIF Framework	4	0.04	0.01	0.02	1.00	0.00
Favored Group x Number of Clusters x R:F Ratio x Impact Presence x ICC Magnitude	1	0.21	0.21	0.37	0.57	0.00
Favored Group x Number of Clusters x R:F Ratio x Impact Presence x Multilevel DIF Framework	2	10.96	5.48	9.57	0.03	0.00
Favored Group x Number of Clusters x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	2	2.33	1.16	2.03	0.25	0.00
Favored Group x Number of Clusters x Impact Presence x ICC Magnitude x Multilevel DIF Framework	2	7.54	3.77	6.58	0.05	0.00
Favored Group x R:F Ratio x Impact Presence x ICC Magnitude x Multilevel DIF Framework	2	0.51	0.25	0.44	0.67	0.00
DIF Magnitude x Number of Clusters x R:F Ratio x	2	352.32	176.16	307.66	0.00	0.00

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Impact Presence x ICC Magnitude						
DIF Magnitude x Number of Clusters x R:F Ratio x Impact Presence x Multilevel DIF Framework	4	3.20	0.80	1.40	0.38	0.00
DIF Magnitude x Number of Clusters x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	4	4.97	1.24	2.17	0.24	0.00
DIF Magnitude x Number of Clusters x Impact Presence x ICC Magnitude x Multilevel DIF Framework	4	649.20	162.30	283.45	0.00	0.00
DIF Magnitude x R:F Ratio x Impact Presence x ICC Magnitude x Multilevel DIF Framework	4	669.34	167.33	292.25	0.00	0.00
Number of Clusters x R:F Ratio x Impact Presence x ICC Magnitude x Multilevel DIF Framework	2	398.16	199.08	347.69	0.00	0.00
Favored Group x DIF Magnitude x Number of Clusters x R:F Ratio x Impact Presence x ICC Magnitude	2	1.17	0.58	1.02	0.44	0.00
Favored Group x DIF Magnitude x Number of Clusters x R:F Ratio x Impact Presence x Multilevel DIF Framework	4	43.93	10.98	19.18	0.01	0.00
Favored Group x DIF Magnitude x Number of Clusters x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	4	12.46	3.12	5.44	0.06	0.00
Favored Group x DIF Magnitude x Number of Clusters x Impact Presence	4	18.99	4.75	8.29	0.03	0.00

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
x ICC Magnitude x Multilevel DIF Framework						
Favored Group x DIF Magnitude x R:F Ratio x Impact Presence x ICC Magnitude x Multilevel DIF Framework	4	0.24	0.06	0.11	0.97	0.00
Favored Group x Number of Clusters x R:F Ratio x Impact Presence x ICC Magnitude x Multilevel DIF Framework	2	0.40	0.20	0.35	0.72	0.00
DIF Magnitude x Number of Clusters x R:F Ratio x Impact Presence x ICC Magnitude x Multilevel DIF Framework	4	701.02	175.26	306.08	0.00	0.00
Error	4	2.29	0.57			
Total	287	55122.25				

APPENDIX B

RELATIVE BIAS RESULTS

Table B.1

BMH95 Bias Results for Invariant Items

		High (ICC = 0.2)				Low (ICC = 0.1)			
		None (Δ=0.0)		Impact (Δ=0.5)		None (Δ=0.0)		Impact (Δ=0.5)	
		Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)
Null (δ=0.0)	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	300	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Small (δ=0.2)	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	300	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Medium (δ=0.4)	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	300	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Large (δ=0.6)	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	300	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table B.2

BMH95 Bias Results for DIF Contaminated Items

		High (ICC = 0.2)				Low (ICC = 0.1)			
		None (Δ=0.0)		Impact (Δ=0.5)		None (Δ=0.0)		Impact (Δ=0.5)	
		Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)
Null									
(δ=0.0)	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	300	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Small									
(δ=0.2)	100	0.000	-0.005	-0.014	-0.011	0.000	0.000	-0.002	-0.008
	300	0.000	0.000	-0.008	-0.009	0.000	0.001	-0.004	-0.010
Medium									
(δ=0.4)	100	0.002	-0.009	-0.011	-0.008	-0.003	-0.003	-0.001	-0.008
	300	-0.001	-0.004	-0.12	-0.009	0.000	0.002	-0.009	-0.008
Large									
(δ=0.6)	100	-0.001	-0.002	-0.006	-0.008	-0.002	-0.001	-0.004	-0.008
	300	0.001	0.000	-0.010	-0.010	0.000	0.001	-0.006	-0.007

Table B.3

Multilevel Rasch Model Bias Results for Invariant Items

		High (ICC = 0.2)				Low (ICC = 0.1)			
		None (Δ=0.0)		Impact (Δ=0.5)		None (Δ=0.0)		Impact (Δ=0.5)	
		Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)
Null (δ=0.0)	100	0.008	-0.012	0.064	0.056	0.000	-0.003	0.028	0.038
	300	0.001	0.001	0.062	0.072	0.004	0.001	0.024	0.024
Small (δ=0.2)	100	-0.007	0.001	0.066	0.073	-0.010	0.007	0.025	0.012
	300	0.000	0.002	0.055	0.068	-0.003	0.006	0.025	0.026
Medium (δ=0.4)	100	0.004	-0.006	0.063	0.057	0.009	-0.012	0.026	0.026
	300	-0.001	-0.009	0.066	0.065	0.004	0.002	0.025	0.027
Large (δ=0.6)	100	0.001	0.000	0.071	0.062	0.003	0.009	0.031	0.034
	300	-0.005	0.003	0.068	0.066	0.000	-0.005	0.026	0.024

Table B.4

Multilevel Rasch Model Bias Results for DIF Contaminated Items

		High (ICC = 0.2)				Low (ICC = 0.1)			
		None (Δ=0.0)		Impact (Δ=0.5)		None (Δ=0.0)		Impact (Δ=0.5)	
		Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)
Null (δ=0.0)	100	0.005	-0.015	-0.035	-0.047	0.004	-0.012	-0.057	-0.046
	300	0.000	0.000	-0.031	-0.033	0.000	-0.005	-0.060	-0.048
Small (δ=0.2)	100	-0.004	-0.010	-0.042	-0.033	-0.009	0.022	-0.048	-0.061
	300	0.000	0.001	-0.040	-0.028	0.000	0.006	-0.054	-0.061
Medium (δ=0.4)	100	0.003	-0.004	-0.036	-0.039	0.001	-0.008	-0.046	-0.065
	300	-0.003	-0.012	-0.037	-0.031	0.005	0.006	-0.059	-0.063
Large (δ=0.6)	100	-0.004	-0.003	-0.022	-0.032	0.008	0.000	-0.048	-0.056
	300	0.001	0.003	-0.027	-0.032	-0.003	-0.003	-0.054	-0.056

Table B.5

SIBTEST BSSE Bias Results for Invariant Items

		High (ICC = 0.2)				Low (ICC = 0.1)			
		None (Δ=0.0)		Impact (Δ=0.5)		None (Δ=0.0)		Impact (Δ=0.5)	
		Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)
Null									
(δ=0.0)	100	0.001	0.000	0.002	0.011	0.003	-0.002	-0.001	0.013
	300	0.000	-0.001	0.003	0.006	0.000	-0.005	0.002	0.002
Small									
(δ=0.2)	100	0.007	0.001	0.006	0.007	-0.001	0.000	0.000	0.007
	300	0.000	0.002	0.001	0.002	-0.001	0.000	0.001	0.000
Medium									
(δ=0.4)	100	-0.002	-0.002	0.010	0.010	0.000	0.000	0.000	0.006
	300	-0.002	0.001	-0.001	-0.001	0.003	0.000	-0.001	0.001
Large									
(δ=0.6)	100	0.006	0.000	0.007	0.004	-0.002	0.003	0.005	0.008
	300	0.001	-0.004	0.001	0.006	0.000	0.000	0.000	0.001

Table B.6

SIBTEST BSSE Bias Results for DIF Contaminated Items

		High (ICC = 0.2)				Low (ICC = 0.1)			
		None (Δ=0.0)		Impact (Δ=0.5)		None (Δ=0.0)		Impact (Δ=0.5)	
		Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)
Null									
(δ=0.0)	100	0.007	0.002	0.006	0.007	0.003	-0.002	0.003	0.006
	300	0.001	-0.002	-0.002	0.004	0.000	-0.003	0.000	0.002
Small									
(δ=0.2)	100	-0.001	0.004	0.005	0.009	0.003	0.004	0.001	0.006
	300	0.001	-0.001	-0.005	0.000	0.001	0.001	0.000	0.001
Medium									
(δ=0.4)	100	-0.002	0.003	0.003	0.007	-0.006	0.000	0.000	0.007
	300	-0.002	0.002	0.001	0.000	0.002	0.003	-0.001	0.003
Large									
(δ=0.6)	100	0.001	0.004	0.004	0.008	-0.001	0.002	0.006	0.008
	300	0.004	0.001	0.001	0.000	0.001	-0.001	-0.003	0.001

Table B.7

The ANOVA Results for Relative Bias

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Item Type	1	0.026	0.026	7247.829	0.000	**0.163**
Magnitude DIF	3	0.000	0.000	13.625	0.004	0.001
Number of Clusters	1	0.000	0.000	2.853	0.142	0.000
R:F Ratio	1	0.000	0.000	1.616	0.251	0.000
Impact Presence	1	0.000	0.000	74.160	0.000	0.002
ICC Magnitude	1	0.002	0.002	561.931	0.000	0.013
Multilevel DIF Framework	2	0.003	0.002	455.223	0.000	0.021
Item Type x Magnitude of DIF	3	0.000	0.000	1.528	0.301	0.000
Item Type x Number of Clusters	1	0.000	0.000	0.421	0.540	0.000
Item Type x R:F Ratio	1	0.000	0.000	0.084	0.781	0.000
Item Type x Impact Presence	1	0.026	0.026	7220.603	0.000	**0.163**
Item Type x ICC Magnitude	1	0.000	0.000	99.070	0.000	0.002
Item Type x Multilevel DIF Framework	2	0.042	0.021	5858.980	0.000	**0.264**
DIF Magnitude x Number of Clusters	3	0.000	0.000	6.358	0.027	0.000
DIF Magnitude x R:F Ratio	3	0.000	0.000	10.373	0.009	0.001
DIF Magnitude x Impact Presence	3	0.000	0.000	4.684	0.052	0.000
DIF Magnitude x ICC Magnitude	3	0.000	0.000	1.540	0.298	0.000
DIF Magnitude x Multilevel DIF Framework	6	0.000	0.000	5.448	0.029	0.001
Number of Clusters x R:F Ratio	1	0.000	0.000	28.966	0.002	0.001
Number of Clusters x Impact Presence	1	0.000	0.000	10.786	0.017	0.000
Number of Clusters x ICC Magnitude	1	0.000	0.000	1.522	0.263	0.000
Number of Clusters x Multilevel DIF Framework	2	0.000	0.000	11.913	0.008	0.001
R:F Ratio x Impact Presence	1	0.000	0.000	31.777	0.001	0.001
R:F Ratio x ICC Magnitude	1	0.000	0.000	5.869	0.052	0.000
R:F Ratio x Multilevel DIF Framework	2	0.000	0.000	57.361	0.000	0.003
Impact Presence x ICC Magnitude	1	0.003	0.003	736.829	0.000	0.017

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Impact Presence x Multilevel DIF Framework	2	0.002	0.001	262.665	0.000	0.012
ICC Magnitude x Multilevel DIF Framework	2	0.004	0.002	579.451	0.000	0.026
Item Type x DIF Magnitude x Number of Clusters	3	0.000	0.000	0.374	0.775	0.000
Item Type x DIF Magnitude x R:F Ratio	3	0.000	0.000	4.048	0.069	0.000
Item Type x DIF Magnitude x Impact Presence	3	0.000	0.000	0.181	0.905	0.000
Item Type x DIF Magnitude x ICC Magnitude	3	0.000	0.000	4.676	0.052	0.000
Item Type x DIF Magnitude x Multilevel DIF Framework	6	0.000	0.000	0.410	0.849	0.000
Item Type x Number of Clusters x R:F Ratio	1	0.000	0.000	1.406	0.280	0.000
Item Type x Number of Clusters x Impact Presence	1	0.000	0.000	2.620	0.157	0.000
Item Type x Number of Clusters x ICC Magnitude	1	0.000	0.000	0.357	0.572	0.000
Item Type x Number of Clusters x Multilevel DIF Framework	2	0.000	0.000	1.650	0.269	0.000
Item Type x R:F Ratio x Impact Presence	1	0.000	0.000	1.194	0.317	0.000
Item Type x R:F Ratio x ICC Magnitude	1	0.000	0.000	0.546	0.488	0.000
Item Type x R:F Ratio x Multilevel DIF Framework	2	0.000	0.000	19.537	0.002	0.001
Item Type x Impact Presence x ICC Magnitude	1	0.000	0.000	84.961	0.000	0.002
Item Type x Impact Presence x Multilevel DIF Framework	2	0.038	0.019	5322.768	0.000	**0.240**
Item Type x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	41.836	0.000	0.002
DIF Magnitude x Number of Clusters x R:F Ratio	3	0.000	0.000	3.239	0.103	0.000
DIF Magnitude x Number of Clusters x Impact Presence	3	0.000	0.000	1.437	0.322	0.000
DIF Magnitude x Number of Clusters x ICC Magnitude	3	0.000	0.000	12.605	0.005	0.001

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
DIF Magnitude x Number of Clusters x Multilevel DIF Framework	6	0.000	0.000	5.827	0.025	0.001
DIF Magnitude x R:F Ratio x Impact Presence	3	0.000	0.000	24.532	0.001	0.002
DIF Magnitude x R:F Ratio x ICC Magnitude	3	0.000	0.000	7.856	0.017	0.001
DIF Magnitude x R:F Ratio x Multilevel DIF Framework	6	0.000	0.000	13.615	0.003	0.002
DIF Magnitude x Impact Presence x ICC Magnitude	3	0.000	0.000	3.155	0.107	0.000
DIF Magnitude x Impact Presence x Multilevel DIF Framework	6	0.000	0.000	1.791	0.248	0.000
DIF Magnitude x ICC Magnitude x Multilevel DIF Framework	6	0.000	0.000	2.096	0.195	0.000
Number of Clusters x R:F Ratio x Impact Presence	1	0.000	0.000	2.586	0.159	0.000
Number of Clusters x R:F Ratio x ICC Magnitude	1	0.000	0.000	5.869	0.052	0.000
Number of Clusters x R:F Ratio x Multilevel DIF Framework	2	0.000	0.000	18.170	0.003	0.001
Number of Clusters x Impact Presence x ICC Magnitude	1	0.000	0.000	6.903	0.039	0.000
Number of Clusters x Impact Presence x Multilevel DIF Framework	2	0.000	0.000	1.290	0.342	0.000
Number of Clusters x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	11.543	0.009	0.001
R:F Ratio x Impact Presence x ICC Magnitude	1	0.000	0.000	3.149	0.126	0.000
R:F Ratio x Impact Presence x Multilevel DIF Framework	2	0.000	0.000	13.879	0.006	0.001
R:F Ratio x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	1.272	0.346	0.000
Impact Presence x ICC Magnitude x Multilevel DIF Framework	2	0.005	0.003	764.354	0.000	0.034
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio	3	0.000	0.000	1.245	0.373	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Item Type x DIF Magnitude x Number of Clusters x Impact Presence	3	0.000	0.000	2.881	0.125	0.000
Item Type x DIF Magnitude x Number of Clusters x ICC Magnitude	3	0.000	0.000	4.073	0.068	0.000
Item Type x DIF Magnitude x Number of Clusters x Multilevel DIF Framework	6	0.000	0.000	1.975	0.214	0.000
Item Type x DIF Magnitude x R:F Ratio x Impact Presence	3	0.000	0.000	1.561	0.294	0.000
Item Type x DIF Magnitude x R:F Ratio x ICC Magnitude	3	0.000	0.000	2.519	0.155	0.000
Item Type x DIF Magnitude x R:F Ratio x Multilevel DIF Framework	6	0.000	0.000	2.748	0.122	0.000
Item Type x DIF Magnitude x Impact Presence x ICC Magnitude	3	0.000	0.000	0.554	0.664	0.000
Item Type x DIF Magnitude x Impact Presence x Multilevel DIF Framework	6	0.000	0.000	0.876	0.562	0.000
Item Type x DIF Magnitude x ICC Magnitude x Multilevel DIF Framework	6	0.000	0.000	0.727	0.646	0.000
Item Type x Number of Clusters x R:F Ratio x Impact Presence	1	0.000	0.000	1.538	0.261	0.000
Item Type x Number of Clusters x R:F Ratio x ICC Magnitude	1	0.000	0.000	10.337	0.018	0.000
Item Type x Number of Clusters x R:F Ratio x Multilevel DIF Framework	2	0.000	0.000	1.527	0.291	0.000
Item Type x Number of Clusters x Impact Presence x ICC Magnitude	1	0.000	0.000	1.466	0.272	0.000
Item Type x Number of Clusters x Impact Presence x Multilevel DIF Framework	2	0.000	0.000	0.045	0.956	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Item Type x Number of Clusters x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	1.285	0.343	0.000
Item Type x R:F Ratio x Impact Presence x ICC Magnitude	1	0.000	0.000	1.474	0.270	0.000
Item Type x R:F Ratio x Impact Presence x Multilevel DIF Framework	2	0.000	0.000	0.080	0.924	0.000
Item Type x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.628	0.566	0.000
Item Type x Impact Presence x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	43.551	0.000	0.002
DIF Magnitude x Number of Clusters x R:F Ratio x Impact Presence	3	0.000	0.000	5.007	0.045	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x ICC Magnitude	3	0.000	0.000	5.711	0.034	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Multilevel DIF Framework	6	0.000	0.000	2.396	0.156	0.000
DIF Magnitude x Number of Clusters x Impact Presence x ICC Magnitude	3	0.000	0.000	7.859	0.017	0.001
DIF Magnitude x Number of Clusters x Impact Presence x Multilevel DIF Framework	6	0.000	0.000	3.412	0.080	0.000
DIF Magnitude x Number of Clusters x ICC Magnitude x Multilevel DIF Framework	6	0.000	0.000	4.913	0.037	0.001
DIF Magnitude x R:F Ratio x Impact Presence x ICC Magnitude	3	0.000	0.000	20.926	0.001	0.001
DIF Magnitude x R:F Ratio x Impact Presence x Multilevel DIF Framework	6	0.000	0.000	10.449	0.006	0.001
DIF Magnitude x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	6	0.000	0.000	5.809	0.025	0.001

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
DIF Magnitude x Impact Presence x ICC Magnitude x Multilevel DIF Framework	6	0.000	0.000	2.792	0.119	0.000
Number of Clusters x R:F Ratio x Impact Presence x ICC Magnitude	1	0.000	0.000	0.050	0.830	0.000
Number of Clusters x R:F Ratio x Impact Presence x Multilevel DIF Framework	2	0.000	0.000	1.534	0.290	0.000
Number of Clusters x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	3.037	0.123	0.000
Number of Clusters x Impact Presence x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	4.255	0.071	0.000
R:F Ratio x Impact Presence x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	6.386	0.033	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x Impact Presence	3	0.000	0.000	1.454	0.318	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x ICC Magnitude	3	0.000	0.000	2.172	0.192	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x Multilevel DIF Framework	6	0.000	0.000	1.428	0.338	0.000
Item Type x DIF Magnitude x Number of Clusters x Impact Presence x ICC Magnitude	3	0.000	0.000	7.545	0.018	0.001
Item Type x DIF Magnitude x Number of Clusters x Impact Presence x Multilevel DIF Framework	6	0.000	0.000	0.990	0.505	0.000
Item Type x DIF Magnitude x Number of Clusters x ICC Magnitude x Multilevel DIF Framework	6	0.000	0.000	1.331	0.369	0.000
Item Type x DIF Magnitude x R:F Ratio x Impact Presence x ICC Magnitude	3	0.000	0.000	7.727	0.017	0.001

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Item Type x DIF Magnitude x R:F Ratio x Impact Presence x Multilevel DIF Framework	6	0.000	0.000	1.132	0.442	0.000
Item Type x DIF Magnitude x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	6	0.000	0.000	1.032	0.485	0.000
Item Type x DIF Magnitude x Impact Presence x ICC Magnitude x Multilevel DIF Framework	6	0.000	0.000	1.792	0.248	0.000
Item Type x Number of Clusters x R:F Ratio x Impact Presence x ICC Magnitude	1	0.000	0.000	0.524	0.496	0.000
Item Type x Number of Clusters x R:F Ratio x Impact Presence x Multilevel DIF Framework	2	0.000	0.000	1.055	0.405	0.000
Item Type x Number of Clusters x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	1.677	0.264	0.000
Item Type x Number of Clusters x Impact Presence x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.490	0.635	0.000
Item Type x R:F Ratio x Impact Presence x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.556	0.601	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Impact Presence x ICC Magnitude	3	0.000	0.000	0.496	0.698	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Impact Presence x Multilevel DIF Framework	6	0.000	0.000	3.550	0.074	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	6	0.000	0.000	5.355	0.030	0.001
DIF Magnitude x Number of Clusters x Impact Presence x	6	0.000	0.000	8.792	0.009	0.001

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
ICC Magnitude x Multilevel DIF Framework						
DIF Magnitude x R:F Ratio x Impact Presence x ICC Magnitude x Multilevel DIF Framework	6	0.000	0.000	12.699	0.003	0.002
Number of Clusters x R:F Ratio x Impact Presence x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.520	0.619	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x Impact Presence x ICC Magnitude	3	0.000	0.000	0.384	0.769	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x Impact Presence x Multilevel DIF Framework	6	0.000	0.000	2.234	0.176	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	6	0.000	0.000	2.087	0.196	0.000
Item Type x DIF Magnitude x Number of Clusters x Impact Presence x ICC Magnitude x Multilevel DIF Framework	6	0.000	0.000	3.014	0.103	0.000
Item Type x DIF Magnitude x R:F Ratio x Impact Presence x ICC Magnitude x Multilevel DIF Framework	6	0.000	0.000	4.666	0.041	0.001
Item Type x Number of Clusters x R:F Ratio x Impact Presence x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.107	0.900	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Impact Presence x ICC Magnitude x Multilevel DIF Framework	6	0.000	0.000	2.206	0.179	0.000
Error	6	0.000	0.000			
Total	383	0.158				

Table B.8

The ANOVA Results for Relative Bias Including Favored Group

		Sum of	Mean			
Source	*df*	Squares	Square	*F*	*p*	η^2
Item Type	2	0.250	0.125	491.728	0.000	0.003
DIF Magnitude	3	0.002	0.001	2.305	0.075	0.000
Number of Clusters	1	0.000	0.000	1.340	0.247	0.000
R:F Ratio	1	0.000	0.000	0.679	0.410	0.000
Impact	1	0.267	0.267	1050.496	0.000	0.009
ICC Magnitude	1	0.056	0.056	220.982	0.000	0.002
Group Favored	2	14.146	7.073	27876.749	0.000	**0.459**
Multilevel DIF Framework	2	0.380	0.190	748.523	0.000	**0.012**
Item Type x DIF Magnitude	5	0.000	0.000	0.258	0.936	0.000
Item Type x Number of Clusters	2	0.000	0.000	0.054	0.947	0.000
Item Type x R:F Ratio	2	0.000	0.000	0.497	0.608	0.000
Item Type x Impact	2	0.250	0.125	493.482	0.000	0.003
Item Type x ICC Magnitude	2	0.006	0.003	11.558	0.000	0.000
Item Type x Group Favored	1	3.525	3.525	13891.453	0.000	0.000
Item Type x Multilevel DIF Framework	1	0.349	0.349	1376.420	0.000	0.003
DIF Magnitude x Number of Clusters	3	0.001	0.000	1.890	0.129	0.000
DIF Magnitude x R:F Ratio	3	0.002	0.001	2.828	0.037	0.000
DIF Magnitude x Impact	3	0.001	0.000	1.223	0.299	0.000
DIF Magnitude x ICC Magnitude	3	0.001	0.000	1.064	0.363	0.000
DIF Magnitude x Group Favored	2	1.069	0.534	2105.746	0.000	**0.035**
DIF Magnitude x Multilevel DIF Framework	6	0.002	0.000	1.232	0.286	0.000
Number of Clusters x R:F Ratio	1	0.001	0.001	5.357	0.021	0.000
Number of Clusters x Impact	1	0.000	0.000	1.365	0.243	0.000
Number of Clusters x ICC Magnitude	1	0.000	0.000	0.087	0.768	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Number of Clusters x Group Favored	2	0.000	0.000	0.129	0.879	0.000
Number of Clusters x Multilevel DIF Framework	2	0.001	0.000	1.930	0.145	0.000
R:F Ratio x Impact	1	0.003	0.003	11.173	0.001	0.000
R:F Ratio x ICC Magnitude	1	0.001	0.001	2.307	0.129	0.000
R:F Ratio x Group Favored	2	0.000	0.000	0.945	0.389	0.000
R:F Ratio x Multilevel DIF Framework	2	0.003	0.001	5.227	0.005	0.000
Impact x ICC Magnitude	1	0.067	0.067	264.550	0.000	0.002
Impact x Group Favored	2	0.000	0.000	0.054	0.947	0.000
Impact x Multilevel DIF Framework	2	0.410	0.205	807.261	0.000	**0.013**
ICC Magnitude x Group Favored	2	0.001	0.000	0.999	0.368	0.000
ICC Magnitude x Multilevel DIF Framework	2	0.103	0.052	203.489	0.000	0.003
Group Favored x Multilevel DIF Framework	2	6.451	3.225	12711.767	0.000	**0.209**
Item Type x DIF Magnitude x Number of Clusters	5	0.000	0.000	0.129	0.986	0.000
Item Type x DIF Magnitude x R:F Ratio	5	0.000	0.000	0.292	0.918	0.000
Item Type x DIF Magnitude x Impact	5	0.000	0.000	0.031	1.000	0.000
Item Type x DIF Magnitude x ICC Magnitude	5	0.000	0.000	0.393	0.854	0.000
Item Type x DIF Magnitude x Group Favored	2	0.426	0.213	839.481	0.000	0.000
Item Type x DIF Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.044	0.957	0.000
Item Type x Number of Clusters x R:F Ratio	2	0.000	0.000	0.202	0.817	0.000
Item Type x Number of Clusters x Impact	2	0.000	0.000	0.185	0.831	0.000
Item Type x Number of Clusters x ICC Magnitude	2	0.000	0.000	0.042	0.959	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Item Type x Number of Clusters x Group Favored	1	0.000	0.000	0.012	0.913	0.000
Item Type x Number of Clusters x Multilevel DIF Framework	1	0.000	0.000	0.455	0.500	0.000
Item Type x R:F Ratio x Impact	2	0.000	0.000	0.379	0.685	0.000
Item Type x R:F Ratio x ICC Magnitude	2	0.000	0.000	0.038	0.962	0.000
Item Type x R:F Ratio x Group Favored	1	0.000	0.000	0.109	0.741	0.000
Item Type x R:F Ratio x Multilevel DIF Framework	1	0.001	0.001	5.055	0.025	0.000
Item Type x Impact x ICC Magnitude	2	0.005	0.003	10.809	0.000	0.000
Item Type x Impact x Group Favored	1	0.000	0.000	0.180	0.671	0.000
Item Type x Impact x Multilevel DIF Framework	1	0.304	0.304	1199.795	0.000	0.002
Item Type x ICC Magnitude x Group Favored	1	0.000	0.000	1.064	0.302	0.000
Item Type x ICC Magnitude x Multilevel DIF Framework	1	0.003	0.003	10.112	0.001	0.000
DIF Magnitude x Number of Clusters x R:F Ratio	3	0.000	0.000	0.354	0.786	0.000
DIF Magnitude x Number of Clusters x Impact	3	0.000	0.000	0.260	0.854	0.000
DIF Magnitude x Number of Clusters x ICC Magnitude	3	0.003	0.001	3.828	0.009	0.000
DIF Magnitude x Number of Clusters x Group Favored	2	0.000	0.000	0.081	0.922	0.000
DIF Magnitude x Number of Clusters x Multilevel DIF Framework	6	0.002	0.000	1.423	0.202	0.000
DIF Magnitude x R:F Ratio x Impact	3	0.005	0.002	6.503	0.000	0.000
DIF Magnitude x R:F Ratio x ICC Magnitude	3	0.001	0.000	1.054	0.368	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
DIF Magnitude x R:F Ratio x Group Favored	2	0.000	0.000	0.009	0.991	0.000
DIF Magnitude x R:F Ratio x Multilevel DIF Framework	6	0.006	0.001	4.246	0.000	0.000
DIF Magnitude x Impact x ICC Magnitude	3	0.001	0.000	0.689	0.559	0.000
DIF Magnitude x Impact x Group Favored	2	0.000	0.000	0.117	0.890	0.000
DIF Magnitude x Impact x Multilevel DIF Framework	6	0.001	0.000	0.343	0.914	0.000
DIF Magnitude x ICC Magnitude x Group Favored	2	0.000	0.000	0.254	0.776	0.000
DIF Magnitude x ICC Magnitude x Multilevel DIF Framework	6	0.001	0.000	0.808	0.564	0.000
DIF Magnitude x Group Favored x Multilevel DIF Framework	2	0.764	0.382	1504.564	0.000	0.000
Number of Clusters x R:F Ratio x Impact	1	0.000	0.000	0.149	0.700	**0.025**
Number of Clusters x R:F Ratio x ICC Magnitude	1	0.001	0.001	5.111	0.024	0.000
Number of Clusters x R:F Ratio x Group Favored	2	0.000	0.000	0.742	0.476	0.000
Number of Clusters x R:F Ratio x Multilevel DIF Framework	2	0.002	0.001	3.328	0.036	0.000
Number of Clusters x Impact x ICC Magnitude	1	0.000	0.000	0.713	0.398	0.000
Number of Clusters x Impact x Group Favored	2	0.000	0.000	0.430	0.651	0.000
Number of Clusters x Impact x Multilevel DIF Framework	2	0.000	0.000	0.314	0.731	0.000
Number of Clusters x ICC Magnitude x Group Favored	2	0.000	0.000	0.297	0.743	0.000
Number of Clusters x ICC Magnitude x Multilevel DIF Framework	2	0.001	0.000	1.949	0.142	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Number of Clusters x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.126	0.882	0.000
R:F Ratio x Impact x ICC Magnitude	1	0.000	0.000	0.144	0.704	0.000
R:F Ratio x Impact x Group Favored	2	0.000	0.000	0.056	0.945	0.000
R:F Ratio x Impact x Multilevel DIF Framework	2	0.002	0.001	3.947	0.019	0.000
R:F Ratio x ICC Magnitude x Group Favored	2	0.000	0.000	0.887	0.412	0.000
R:F Ratio x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.687	0.503	0.000
R:F Ratio x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.934	0.393	0.000
Impact x ICC Magnitude x Group Favored	2	0.000	0.000	0.396	0.673	0.000
Impact x ICC Magnitude x Multilevel DIF Framework	2	0.133	0.067	262.491	0.000	0.000
Impact x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.548	0.578	0.004
ICC Magnitude x Group Favored x Multilevel DIF Framework	2	0.019	0.010	37.643	0.000	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio	5	0.000	0.000	0.040	0.999	0.001
Item Type x DIF Magnitude x Number of Clusters x Impact	5	0.000	0.000	0.077	0.996	0.000
Item Type x DIF Magnitude x Number of Clusters x ICC Magnitude	5	0.001	0.000	0.398	0.850	0.000
Item Type x DIF Magnitude x Number of Clusters x Group Favored	2	0.000	0.000	0.023	0.978	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Item Type x DIF Magnitude x Number of Clusters x Multilevel DIF Framework	2	0.000	0.000	0.647	0.524	0.000
Item Type x DIF Magnitude x R:F Ratio x Impact	5	0.000	0.000	0.297	0.915	0.000
Item Type x DIF Magnitude x R:F Ratio x ICC Magnitude	5	0.000	0.000	0.006	1.000	0.000
Item Type x DIF Magnitude x R:F Ratio x Group Favored	2	0.000	0.000	0.020	0.980	0.000
Item Type x DIF Magnitude x R:F Ratio x Multilevel DIF Framework	2	0.000	0.000	0.148	0.862	0.000
Item Type x DIF Magnitude x Impact x ICC Magnitude	5	0.000	0.000	0.149	0.980	0.000
Item Type x DIF Magnitude x Impact x Group Favored	2	0.000	0.000	0.101	0.904	0.000
Item Type x DIF Magnitude x Impact x Multilevel DIF Framework	2	0.000	0.000	0.279	0.756	0.000
Item Type x DIF Magnitude x ICC Magnitude x Group Favored	2	0.000	0.000	0.249	0.780	0.000
Item Type x DIF Magnitude x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.127	0.881	0.000
Item Type x Number of Clusters x R:F Ratio x Impact	2	0.000	0.000	0.182	0.833	0.000
Item Type x Number of Clusters x R:F Ratio x ICC Magnitude	2	0.000	0.000	0.667	0.513	0.000
Item Type x Number of Clusters x R:F Ratio x Group Favored	1	0.000	0.000	0.015	0.903	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Item Type x Number of Clusters x R:F Ratio x Multilevel DIF Framework	1	0.000	0.000	0.443	0.506	0.000
Item Type x Number of Clusters x Impact x ICC Magnitude	2	0.000	0.000	0.128	0.880	0.000
Item Type x Number of Clusters x Impact x Group Favored	1	0.000	0.000	0.063	0.801	0.000
Item Type x Number of Clusters x Impact x Multilevel DIF Framework	1	0.000	0.000	0.012	0.912	0.000
Item Type x Number of Clusters x ICC Magnitude x Group Favored	1	0.000	0.000	0.008	0.929	0.000
Item Type x Number of Clusters x ICC Magnitude x Multilevel DIF Framework	1	0.000	0.000	0.329	0.567	0.000
Item Type x R:F Ratio x Impact x ICC Magnitude	2	0.000	0.000	0.173	0.841	0.000
Item Type x R:F Ratio x Impact x Group Favored	1	0.000	0.000	0.039	0.844	0.000
Item Type x R:F Ratio x Impact x Multilevel DIF Framework	1	0.000	0.000	0.000	0.999	0.000
Item Type x R:F Ratio x ICC Magnitude x Group Favored	1	0.000	0.000	0.073	0.788	0.000
Item Type x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	1	0.000	0.000	0.005	0.943	0.000
Item Type x Impact x ICC Magnitude x Group Favored	1	0.000	0.000	0.001	0.973	0.000
Item Type x Impact x ICC Magnitude x Multilevel DIF Framework	1	0.002	0.002	9.831	0.002	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Impact	3	0.001	0.000	1.573	0.194	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
DIF Magnitude x Number of Clusters x R:F Ratio x ICC Magnitude	3	0.001	0.000	1.883	0.130	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Group Favored	2	0.000	0.000	0.283	0.754	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Multilevel DIF Framework	6	0.002	0.000	1.268	0.268	0.000
DIF Magnitude x Number of Clusters x Impact x ICC Magnitude	3	0.001	0.000	1.269	0.283	0.000
DIF Magnitude x Number of Clusters x Impact x Group Favored	2	0.000	0.000	0.577	0.561	0.000
DIF Magnitude x Number of Clusters x Impact x Multilevel DIF Framework	6	0.002	0.000	1.423	0.201	0.000
DIF Magnitude x Number of Clusters x ICC Magnitude x Group Favored	2	0.001	0.000	1.670	0.188	0.000
DIF Magnitude x Number of Clusters x ICC Magnitude x Multilevel DIF Framework	6	0.003	0.000	1.769	0.101	0.000
DIF Magnitude x Number of Clusters x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.183	0.833	0.000
DIF Magnitude x R:F Ratio x Impact x ICC Magnitude	3	0.003	0.001	3.429	0.016	0.000
DIF Magnitude x R:F Ratio x Impact x Group Favored	2	0.000	0.000	0.484	0.617	0.000
DIF Magnitude x R:F Ratio x Impact x Multilevel DIF Framework	6	0.005	0.001	3.297	0.003	0.000
DIF Magnitude x R:F Ratio x ICC Magnitude x Group Favored	2	0.000	0.000	0.021	0.979	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
DIF Magnitude x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	6	0.002	0.000	1.601	0.142	0.000
DIF Magnitude x R:F Ratio x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.118	0.889	0.000
DIF Magnitude x Impact x ICC Magnitude x Group Favored	2	0.000	0.000	0.111	0.895	0.000
DIF Magnitude x Impact x ICC Magnitude x Multilevel DIF Framework	6	0.001	0.000	0.544	0.775	0.000
DIF Magnitude x Impact x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.211	0.810	0.000
DIF Magnitude x ICC Magnitude x Group Favored x Multilevel DIF Framework	2	0.001	0.001	2.195	0.111	0.000
Number of Clusters x R:F Ratio x Impact x ICC Magnitude	1	0.000	0.000	0.014	0.905	0.000
Number of Clusters x R:F Ratio x Impact x Group Favored	2	0.000	0.000	0.058	0.944	0.000
Number of Clusters x R:F Ratio x Impact x Multilevel DIF Framework	2	0.000	0.000	0.048	0.953	0.000
Number of Clusters x R:F Ratio x ICC Magnitude x Group Favored	2	0.000	0.000	0.499	0.607	0.000
Number of Clusters x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	2	0.001	0.001	1.986	0.137	0.000
Number of Clusters x R:F Ratio x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.514	0.598	0.000
Number of Clusters x Impact x ICC Magnitude x Group Favored	2	0.000	0.000	0.160	0.852	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Number of Clusters x Impact x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.775	0.461	0.000
Number of Clusters x Impact x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.589	0.555	0.000
Number of Clusters x ICC Magnitude x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.203	0.817	0.000
R:F Ratio x Impact x ICC Magnitude x Group Favored	2	0.001	0.000	1.195	0.303	0.000
R:F Ratio x Impact x ICC Magnitude x Multilevel DIF Framework	2	0.001	0.000	1.450	0.235	0.000
R:F Ratio x Impact x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.150	0.860	0.000
R:F Ratio x ICC Magnitude x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.087	0.916	0.000
Impact x ICC Magnitude x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.480	0.619	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x Impact	5	0.000	0.000	0.140	0.983	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x ICC Magnitude	5	0.000	0.000	0.215	0.956	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x Group Favored	2	0.000	0.000	0.037	0.964	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x Multilevel DIF Framework	2	0.000	0.000	0.229	0.795	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Item Type x DIF Magnitude x Number of Clusters x Impact x ICC Magnitude	5	0.001	0.000	0.718	0.610	0.000
Item Type x DIF Magnitude x Number of Clusters x Impact x Group Favored	2	0.000	0.000	0.024	0.976	0.000
Item Type x DIF Magnitude x Number of Clusters x Impact x Multilevel DIF Framework	2	0.000	0.000	0.136	0.873	0.000
Item Type x DIF Magnitude x Number of Clusters x ICC Magnitude x Group Favored	2	0.000	0.000	0.074	0.928	0.000
Item Type x DIF Magnitude x Number of Clusters x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.428	0.652	0.000
Item Type x DIF Magnitude x R:F Ratio x Impact x ICC Magnitude	5	0.000	0.000	0.285	0.922	0.000
Item Type x DIF Magnitude x R:F Ratio x Impact x Group Favored	2	0.000	0.000	0.045	0.956	0.000
Item Type x DIF Magnitude x R:F Ratio x Impact x Multilevel DIF Framework	2	0.000	0.000	0.141	0.868	0.000
Item Type x DIF Magnitude x R:F Ratio x ICC Magnitude x Group Favored	2	0.000	0.000	0.003	0.997	0.000
Item Type x DIF Magnitude x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.309	0.734	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Item Type x DIF Magnitude x Impact x ICC Magnitude x Group Favored	2	0.000	0.000	0.001	0.999	0.000
Item Type x DIF Magnitude x Impact x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.554	0.575	0.000
Item Type x Number of Clusters x R:F Ratio x Impact x ICC Magnitude	2	0.000	0.000	0.074	0.929	0.000
Item Type x Number of Clusters x R:F Ratio x Impact x Group Favored	1	0.000	0.000	0.023	0.878	0.000
Item Type x Number of Clusters x R:F Ratio x Impact x Multilevel DIF Framework	1	0.000	0.000	0.223	0.637	0.000
Item Type x Number of Clusters x R:F Ratio x ICC Magnitude x Group Favored	1	0.000	0.000	0.000	0.994	0.000
Item Type x Number of Clusters x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	1	0.000	0.000	0.109	0.741	0.000
Item Type x Number of Clusters x Impact x ICC Magnitude x Group Favored	1	0.000	0.000	0.011	0.917	0.000
Item Type x Number of Clusters x Impact x ICC Magnitude x Multilevel DIF Framework	1	0.000	0.000	0.114	0.735	0.000
Item Type x R:F Ratio x Impact x ICC Magnitude x Group Favored	1	0.000	0.000	0.023	0.880	0.000
Item Type x R:F Ratio x Impact x ICC Magnitude x Multilevel DIF Framework	1	0.000	0.000	0.024	0.876	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
DIF Magnitude x Number of Clusters x R:F Ratio x Impact x ICC Magnitude	3	0.000	0.000	0.255	0.858	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Impact x Group Favored	2	0.000	0.000	0.232	0.793	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Impact x Multilevel DIF Framework	6	0.003	0.000	1.727	0.110	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x ICC Magnitude x Group Favored	2	0.000	0.000	0.089	0.915	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	6	0.004	0.001	2.518	0.020	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.420	0.657	0.000
DIF Magnitude x Number of Clusters x Impact x ICC Magnitude x Group Favored	2	0.000	0.000	0.595	0.551	0.000
DIF Magnitude x Number of Clusters x Impact x ICC Magnitude x Multilevel DIF Framework	6	0.005	0.001	2.990	0.006	0.000
DIF Magnitude x Number of Clusters x Impact x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.017	0.983	0.000
DIF Magnitude x Number of Clusters x ICC Magnitude x Group Favored x Multilevel DIF Framework	2	0.001	0.000	1.344	0.261	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
DIF Magnitude x R:F Ratio x Impact x ICC Magnitude x Group Favored	2	0.000	0.000	0.028	0.973	0.000
DIF Magnitude x R:F Ratio x Impact x ICC Magnitude x Multilevel DIF Framework	6	0.005	0.001	3.000	0.006	0.000
DIF Magnitude x R:F Ratio x Impact x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.208	0.812	0.000
DIF Magnitude x R:F Ratio x ICC Magnitude x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.009	0.991	0.000
DIF Magnitude x Impact x ICC Magnitude x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.045	0.956	0.000
Number of Clusters x R:F Ratio x Impact x ICC Magnitude x Group Favored	2	0.000	0.000	0.311	0.733	0.000
Number of Clusters x R:F Ratio x Impact x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.045	0.956	0.000
Number of Clusters x R:F Ratio x Impact x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.130	0.878	0.000
Number of Clusters x R:F Ratio x ICC Magnitude x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.551	0.576	0.000
Number of Clusters x Impact x ICC Magnitude x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.241	0.786	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
R:F Ratio x Impact x ICC Magnitude x Group Favored x Multilevel DIF Framework	2	0.001	0.000	1.661	0.190	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x Impact x ICC Magnitude	5	0.000	0.000	0.033	0.999	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x Impact x Group Favored	2	0.000	0.000	0.019	0.981	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x Impact x Multilevel DIF Framework	2	0.000	0.000	0.829	0.436	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x ICC Magnitude x Group Favored	2	0.000	0.000	0.032	0.969	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.297	0.743	0.000
Item Type x DIF Magnitude x Number of Clusters x Impact x ICC Magnitude x Group Favored	2	0.000	0.000	0.008	0.992	0.000
Item Type x DIF Magnitude x Number of Clusters x Impact x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.938	0.392	0.000
Item Type x DIF Magnitude x R:F Ratio x Impact x ICC Magnitude x Group Favored	2	0.000	0.000	0.053	0.949	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Item Type x DIF Magnitude x R:F Ratio x Impact x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.175	0.839	0.000
Item Type x Number of Clusters x R:F Ratio x Impact x ICC Magnitude x Group Favored	1	0.000	0.000	0.015	0.903	0.000
Item Type x Number of Clusters x R:F Ratio x Impact x ICC Magnitude x Multilevel DIF Framework	1	0.000	0.000	0.000	0.998	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Impact x ICC Magnitude x Group Favored	2	0.001	0.000	1.306	0.271	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Impact x ICC Magnitude x Multilevel DIF Framework	6	0.001	0.000	0.641	0.697	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Impact x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.285	0.752	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x ICC Magnitude x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.387	0.679	0.000
DIF Magnitude x Number of Clusters x Impact x ICC Magnitude x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.128	0.880	0.000
DIF Magnitude x R:F Ratio x Impact x ICC Magnitude x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.393	0.675	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Number of Clusters x R:F Ratio x Impact x ICC Magnitude x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.038	0.963	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x Impact x ICC Magnitude x Group Favored	2	0.000	0.000	0.002	0.998	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x Impact x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.350	0.705	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Impact x ICC Magnitude x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.973	0.378	0.000
Residuals	6896	1.750	0.000			
Total	7423	30.844				

APPENDIX C

ROOT MEAN SQUARE ERROR RESULTS

Table C.1

BMH95 RMSE Results for Invariant Items

		High (ICC = 0.2)				Low (ICC = 0.1)			
		None (Δ=0.0)		Impact (Δ=0.5)		None (Δ=0.0)		Impact (Δ=0.5)	
		Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)
Null (δ=0.0)	100	0.070	0.081	0.072	0.083	0.067	0.078	0.070	0.080
	300	0.040	0.046	0.041	0.047	0.039	0.045	0.039	0.046
Small (δ=0.2)	100	0.070	0.083	0.070	0.083	0.069	0.080	0.068	0.082
	300	0.040	0.047	0.041	0.048	0.039	0.045	0.039	0.045
Medium (δ=0.4)	100	0.071	0.082	0.073	0.083	0.068	0.078	0.069	0.078
	300	0.040	0.046	0.041	0.049	0.039	0.045	0.040	0.046
Large (δ=0.6)	100	0.071	0.042	0.072	0.084	0.068	0.039	0.070	0.079
	300	0.040	0.031	0.042	0.049	0.038	0.030	0.039	0.045

Table C.2

BMH95 RMSE Results for DIF Contaminated Items

		High (ICC = 0.2)				Low (ICC = 0.1)			
		None (Δ=0.0)		Impact (Δ=0.5)		None (Δ=0.0)		Impact (Δ=0.5)	
		Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)
Null (δ=0.0)	100	0.067	0.078	0.067	0.079	0.062	0.073	0.063	0.076
	300	0.037	0.044	0.040	0.043	0.037	0.042	0.038	0.041
Small (δ=0.2)	100	0.148	0.150	0.149	0.155	0.148	0.158	0.154	0.155
	300	0.141	0.144	0.141	0.137	0.141	0.140	0.138	0.141
Medium (δ=0.4)	100	0.238	0.238	0.233	0.238	0.232	0.231	0.235	0.235
	300	0.226	0.227	0.228	0.228	0.226	0.232	0.228	0.231
Large (δ=0.6)	100	0.305	0.303	0.307	0.305	0.307	0.303	0.303	0.308
	300	0.301	0.299	0.298	0.299	0.297	0.302	0.299	0.300

Table C.3

Multilevel Rasch Model RMSE Results for Invariant Items

		High (ICC = 0.2)				Low (ICC = 0.1)			
		None (Δ=0.0)		Impact (Δ=0.5)		None (Δ=0.0)		Impact (Δ=0.5)	
		Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)
Null (δ=0.0)	100	0.124	0.146	0.133	0.151	0.124	0.140	0.142	0.152
	300	0.073	0.084	0.077	0.088	0.069	0.082	0.093	0.108
Small (δ=0.2)	100	0.131	0.155	0.129	0.148	0.123	0.141	0.136	0.164
	300	0.073	0.084	0.077	0.093	0.070	0.082	0.089	0.106
Medium (δ=0.4)	100	0.131	0.148	0.133	0.151	0.122	0.140	0.141	0.148
	300	0.072	0.086	0.079	0.093	0.069	0.081	0.098	0.104
Large (δ=0.6)	100	0.128	0.145	0.133	0.155	0.120	0.138	0.145	0.159
	300	0.073	0.083	0.080	0.091	0.069	0.082	0.098	0.109

Table C.4

Multilevel Rasch Model RMSE Results for DIF Contaminated Items

		High (ICC = 0.2)				Low (ICC = 0.1)			
		None (Δ=0.0)		Impact (Δ=0.5)		None (Δ=0.0)		Impact (Δ=0.5)	
		Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)
Null									
(δ=0.0)	100	0.138	0.157	0.148	0.167	0.122	0.144	0.128	0.154
	300	0.077	0.093	0.097	0.101	0.072	0.082	0.081	0.087
Small									
(δ=0.2)	100	0.137	0.154	0.144	0.167	0.127	0.146	0.126	0.151
	300	0.072	0.092	0.096	0.108	0.072	0.087	0.080	0.090
Medium									
(δ=0.4)	100	0.141	0.156	0.140	0.173	0.124	0.148	0.134	0.150
	300	0.083	0.091	0.102	0.114	0.070	0.086	0.081	0.087
Large									
(δ=0.6)	100	0.150	0.173	0.164	0.184	0.137	0.158	0.137	0.166
	300	0.108	0.116	0.122	0.136	0.097	0.108	0.095	0.111

Table C.5

SIBTEST BSSE RMSE Results for Invariant Items

		High (ICC = 0.2)				Low (ICC = 0.1)			
		None (Δ=0.0)		Impact (Δ=0.5)		None (Δ=0.0)		Impact (Δ=0.5)	
		Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)
Null (δ=0.0)	100	0.074	0.089	0.074	0.093	0.073	0.084	0.074	0.090
	300	0.040	0.048	0.041	0.049	0.040	0.047	0.041	0.050
Small (δ=0.2)	100	0.072	0.088	0.074	0.091	0.074	0.088	0.076	0.090
	300	0.040	0.048	0.041	0.049	0.041	0.050	0.042	0.049
Medium (δ=0.4)	100	0.075	0.087	0.076	0.090	0.072	0.087	0.076	0.091
	300	0.040	0.046	0.040	0.049	0.041	0.047	0.042	0.049
Large (δ=0.6)	100	0.074	0.085	0.076	0.090	0.074	0.083	0.074	0.089
	300	0.040	0.048	0.041	0.481	0.040	0.047	0.042	0.050

Table C.6

SIBTEST BSSE RMSE Results for DIF Contaminated Items

		High (ICC = 0.2)				Low (ICC = 0.1)			
		None (Δ=0.0)		Impact (Δ=0.5)		None (Δ=0.0)		Impact (Δ=0.5)	
		Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)	Even (1:1)	Uneven (3:1)
Null (δ=0.0)	100	0.074	0.086	0.073	0.087	0.072	0.083	0.075	0.091
	300	0.040	0.048	0.041	0.047	0.041	0.047	0.039	0.049
Small (δ=0.2)	100	0.142	0.151	0.145	0.151	0.136	0.144	0.138	0.146
	300	0.132	0.134	0.132	0.133	0.127	0.130	0.125	0.131
Medium (δ=0.4)	100	0.225	0.225	0.217	0.219	0.211	0.215	0.207	0.213
	300	0.215	0.216	0.215	0.212	0.206	0.205	0.208	0.205
Large (δ=0.6)	100	0.289	0.279	0.278	0.284	0.280	0.280	0.273	0.276
	300	0.287	0.286	0.282	0.287	0.272	0.275	0.273	0.275

Table C.7

The ANOVA Results for Root Mean Square Error

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Item Type	1	0.609	0.609	178079.600	0.000	**0.308**
Magnitude DIF	3	0.339	0.113	33079.710	0.000	**0.171**
Number of Clusters	1	0.116	0.116	33790.080	0.000	**0.058**
R:F Ratio	1	0.011	0.011	3081.289	0.000	0.005
Impact Presence	1	0.002	0.002	521.871	0.000	0.001
ICC Magnitude	1	0.001	0.001	331.522	0.000	0.001
Multilevel DIF Framework	2	0.002	0.001	233.323	0.000	0.001
Item Type x Magnitude of DIF	3	0.337	0.112	32826.290	0.000	**0.170**
Item Type x Number of Clusters	1	0.004	0.004	1138.409	0.000	0.002
Item Type x R:F Ratio	1	0.000	0.000	95.424	0.000	0.000
Item Type x Impact Presence	1	0.000	0.000	50.876	0.000	0.000
Item Type x ICC Magnitude	1	0.001	0.001	319.167	0.000	0.001
Item Type x Multilevel DIF Framework	2	0.248	0.124	36191.120	0.000	**0.125**
DIF Magnitude x Number of Clusters	3	0.002	0.001	157.756	0.000	0.001
DIF Magnitude x R:F Ratio	3	0.000	0.000	16.431	0.003	0.000
DIF Magnitude x Impact Presence	3	0.000	0.000	2.211	0.188	0.000
DIF Magnitude x ICC Magnitude	3	0.000	0.000	6.400	0.027	0.000
DIF Magnitude x Multilevel DIF Framework	6	0.136	0.023	6624.751	0.000	**0.069**
Number of Clusters x R:F Ratio	1	0.001	0.001	214.456	0.000	0.000
Number of Clusters x Impact Presence	1	0.000	0.000	5.807	0.053	0.000
Number of Clusters x ICC Magnitude	1	0.000	0.000	26.295	0.002	0.000
Number of Clusters x Multilevel DIF Framework	2	0.019	0.010	2823.068	0.000	0.010
R:F Ratio x Impact Presence	1	0.000	0.000	5.193	0.063	0.000
R:F Ratio x ICC Magnitude	1	0.000	0.000	0.001	0.974	0.000
R:F Ratio x Multilevel DIF Framework	2	0.002	0.001	222.300	0.000	0.001

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Impact Presence x ICC Magnitude	1	0.000	0.000	23.985	0.003	0.000
Impact Presence x Multilevel DIF Framework	2	0.002	0.001	332.287	0.000	0.001
ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	43.127	0.000	0.000
Item Type x DIF Magnitude x Number of Clusters	3	0.002	0.001	150.961	0.000	0.001
Item Type x DIF Magnitude x R:F Ratio	3	0.000	0.000	8.475	0.014	0.000
Item Type x DIF Magnitude x Impact Presence	3	0.000	0.000	4.641	0.053	0.000
Item Type x DIF Magnitude x ICC Magnitude	3	0.000	0.000	2.633	0.144	0.000
Item Type x DIF Magnitude x Multilevel DIF Framework	6	0.138	0.023	6745.062	0.000	**0.070**
Item Type x Number of Clusters x R:F Ratio	1	0.000	0.000	1.142	0.326	0.000
Item Type x Number of Clusters x Impact Presence	1	0.000	0.000	0.600	0.468	0.000
Item Type x Number of Clusters x ICC Magnitude	1	0.000	0.000	7.587	0.033	0.000
Item Type x Number of Clusters x Multilevel DIF Framework	2	0.002	0.001	268.157	0.000	0.001
Item Type x R:F Ratio x Impact Presence	1	0.000	0.000	0.264	0.626	0.000
Item Type x R:F Ratio x ICC Magnitude	1	0.000	0.000	9.018	0.024	0.000
Item Type x R:F Ratio x Multilevel DIF Framework	2	0.000	0.000	49.055	0.000	0.000
Item Type x Impact Presence x ICC Magnitude	1	0.000	0.000	97.243	0.000	0.000
Item Type x Impact Presence x Multilevel DIF Framework	2	0.000	0.000	6.687	0.030	0.000
Item Type x ICC Magnitude x Multilevel DIF Framework	2	0.002	0.001	239.277	0.000	0.001
DIF Magnitude x Number of Clusters x R:F Ratio	3	0.000	0.000	2.439	0.162	0.000
DIF Magnitude x Number of Clusters x Impact Presence	3	0.000	0.000	2.876	0.125	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
DIF Magnitude x Number of Clusters x ICC Magnitude	3	0.000	0.000	3.057	0.113	0.000
DIF Magnitude x Number of Clusters x Multilevel DIF Framework	6	0.000	0.000	18.380	0.001	0.000
DIF Magnitude x R:F Ratio x Impact Presence	3	0.000	0.000	2.278	0.180	0.000
DIF Magnitude x R:F Ratio x ICC Magnitude	3	0.000	0.000	1.297	0.358	0.000
DIF Magnitude x R:F Ratio x Multilevel DIF Framework	6	0.000	0.000	8.054	0.011	0.000
DIF Magnitude x Impact Presence x ICC Magnitude	3	0.000	0.000	0.818	0.530	0.000
DIF Magnitude x Impact Presence x Multilevel DIF Framework	6	0.000	0.000	2.375	0.158	0.000
DIF Magnitude x ICC Magnitude x Multilevel DIF Framework	6	0.000	0.000	2.061	0.200	0.000
Number of Clusters x R:F Ratio x Impact Presence	1	0.000	0.000	3.249	0.122	0.000
Number of Clusters x R:F Ratio x ICC Magnitude	1	0.000	0.000	2.190	0.189	0.000
Number of Clusters x R:F Ratio x Multilevel DIF Framework	2	0.000	0.000	14.986	0.005	0.000
Number of Clusters x Impact Presence x ICC Magnitude	1	0.000	0.000	1.090	0.337	0.000
Number of Clusters x Impact Presence x Multilevel DIF Framework	2	0.000	0.000	18.863	0.003	0.000
Number of Clusters x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	5.481	0.044	0.000
R:F Ratio x Impact Presence x ICC Magnitude	1	0.000	0.000	1.233	0.309	0.000
R:F Ratio x Impact Presence x Multilevel DIF Framework	2	0.000	0.000	2.677	0.148	0.000
R:F Ratio x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.047	0.954	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Impact Presence x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	5.237	0.048	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio	3	0.000	0.000	1.845	0.240	0.000
Item Type x DIF Magnitude x Number of Clusters x Impact Presence	3	0.000	0.000	0.771	0.551	0.000
Item Type x DIF Magnitude x Number of Clusters x ICC Magnitude	3	0.000	0.000	0.666	0.603	0.000
Item Type x DIF Magnitude x Number of Clusters x Multilevel DIF Framework	6	0.000	0.000	19.991	0.001	0.000
Item Type x DIF Magnitude x R:F Ratio x Impact Presence	3	0.000	0.000	1.996	0.216	0.000
Item Type x DIF Magnitude x R:F Ratio x ICC Magnitude	3	0.000	0.000	0.175	0.910	0.000
Item Type x DIF Magnitude x R:F Ratio x Multilevel DIF Framework	6	0.000	0.000	5.490	0.029	0.000
Item Type x DIF Magnitude x Impact Presence x ICC Magnitude	3	0.000	0.000	1.748	0.256	0.000
Item Type x DIF Magnitude x Impact Presence x Multilevel DIF Framework	6	0.000	0.000	2.475	0.147	0.000
Item Type x DIF Magnitude x ICC Magnitude x Multilevel DIF Framework	6	0.000	0.000	5.735	0.026	0.000
Item Type x Number of Clusters x R:F Ratio x Impact Presence	1	0.000	0.000	4.192	0.087	0.000
Item Type x Number of Clusters x R:F Ratio x ICC Magnitude	1	0.000	0.000	0.078	0.790	0.000
Item Type x Number of Clusters x R:F Ratio x Multilevel DIF Framework	2	0.000	0.000	10.956	0.010	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Item Type x Number of Clusters x Impact Presence x ICC Magnitude	1	0.000	0.000	8.248	0.028	0.000
Item Type x Number of Clusters x Impact Presence x Multilevel DIF Framework	2	0.000	0.000	1.258	0.350	0.000
Item Type x Number of Clusters x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	7.391	0.024	0.000
Item Type x R:F Ratio x Impact Presence x ICC Magnitude	1	0.000	0.000	0.006	0.942	0.000
Item Type x R:F Ratio x Impact Presence x Multilevel DIF Framework	2	0.000	0.000	0.400	0.687	0.000
Item Type x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	1.214	0.361	0.000
Item Type x Impact Presence x ICC Magnitude x Multilevel DIF Framework	2	0.001	0.000	130.988	0.000	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Impact Presence	3	0.000	0.000	0.262	0.851	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x ICC Magnitude	3	0.000	0.000	0.602	0.637	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Multilevel DIF Framework	6	0.000	0.000	1.479	0.323	0.000
DIF Magnitude x Number of Clusters x Impact Presence x ICC Magnitude	3	0.000	0.000	1.346	0.345	0.000
DIF Magnitude x Number of Clusters x Impact Presence x Multilevel DIF Framework	6	0.000	0.000	1.642	0.281	0.000
DIF Magnitude x Number of Clusters x ICC Magnitude x Multilevel DIF Framework	6	0.000	0.000	1.644	0.281	0.000
DIF Magnitude x R:F Ratio x Impact Presence x ICC Magnitude	3	0.000	0.000	4.451	0.057	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
DIF Magnitude x R:F Ratio x Impact Presence x Multilevel DIF Framework	6	0.000	0.000	2.495	0.145	0.000
DIF Magnitude x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	6	0.000	0.000	1.336	0.367	0.000
DIF Magnitude x Impact Presence x ICC Magnitude x Multilevel DIF Framework	6	0.000	0.000	0.317	0.906	0.000
Number of Clusters x R:F Ratio x Impact Presence x ICC Magnitude	1	0.000	0.000	0.579	0.476	0.000
Number of Clusters x R:F Ratio x Impact Presence x Multilevel DIF Framework	2	0.000	0.000	0.192	0.830	0.000
Number of Clusters x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.031	0.970	0.000
Number of Clusters x Impact Presence x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.253	0.785	0.000
R:F Ratio x Impact Presence x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	1.783	0.247	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x Impact Presence	3	0.000	0.000	1.735	0.259	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x ICC Magnitude	3	0.000	0.000	3.542	0.088	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x Multilevel DIF Framework	6	0.000	0.000	1.181	0.422	0.000
Item Type x DIF Magnitude x Number of Clusters x Impact Presence x ICC Magnitude	3	0.000	0.000	3.105	0.110	0.000
Item Type x DIF Magnitude x Number of Clusters x Impact Presence x Multilevel DIF Framework	6	0.000	0.000	0.618	0.713	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Item Type x DIF Magnitude x Number of Clusters x ICC Magnitude x Multilevel DIF Framework	6	0.000	0.000	2.911	0.110	0.000
Item Type x DIF Magnitude x R:F Ratio x Impact Presence x ICC Magnitude	3	0.000	0.000	0.938	0.479	0.000
Item Type x DIF Magnitude x R:F Ratio x Impact Presence x Multilevel DIF Framework	6	0.000	0.000	0.796	0.606	0.000
Item Type x DIF Magnitude x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	6	0.000	0.000	0.339	0.893	0.000
Item Type x DIF Magnitude x Impact Presence x ICC Magnitude x Multilevel DIF Framework	6	0.000	0.000	0.782	0.614	0.000
Item Type x Number of Clusters x R:F Ratio x Impact Presence x ICC Magnitude	1	0.000	0.000	0.520	0.498	0.000
Item Type x Number of Clusters x R:F Ratio x Impact Presence x Multilevel DIF Framework	2	0.000	0.000	1.804	0.243	0.000
Item Type x Number of Clusters x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.948	0.439	0.000
Item Type x Number of Clusters x Impact Presence x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	7.523	0.023	0.000
Item Type x R:F Ratio x Impact Presence x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.147	0.866	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Impact Presence x ICC Magnitude	3	0.000	0.000	0.930	0.482	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Impact	6	0.000	0.000	0.913	0.543	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Presence x Multilevel DIF Framework						
DIF Magnitude x Number of Clusters x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	6	0.000	0.000	1.223	0.407	0.000
DIF Magnitude x Number of Clusters x Impact Presence x ICC Magnitude x Multilevel DIF Framework	6	0.000	0.000	0.931	0.534	0.000
DIF Magnitude x R:F Ratio x Impact Presence x ICC Magnitude x Multilevel DIF Framework	6	0.000	0.000	3.505	0.076	0.000
Number of Clusters x R:F Ratio x Impact Presence x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.130	0.881	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x Impact Presence x ICC Magnitude	3	0.000	0.000	1.687	0.268	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x Impact Presence x Multilevel DIF Framework	6	0.000	0.000	1.000	0.500	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	6	0.000	0.000	0.877	0.561	0.000
Item Type x DIF Magnitude x Number of Clusters x Impact Presence x ICC Magnitude x Multilevel DIF Framework	6	0.000	0.000	1.228	0.405	0.000
Item Type x DIF Magnitude x R:F Ratio x Impact Presence x ICC Magnitude x Multilevel DIF Framework	6	0.000	0.000	3.405	0.081	0.000
Item Type x Number of Clusters x R:F Ratio x Impact	2	0.000	0.000	0.506	0.627	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Presence x ICC Magnitude x Multilevel DIF Framework						
DIF Magnitude x Number of Clusters x R:F Ratio x Impact Presence x ICC Magnitude x Multilevel DIF Framework	6	0.000	0.000	1.389	0.350	0.000
Error	6	0.000	0.000			
Total	383	1.980				

Table C.8

The ANOVA Results for Root Mean Square Error Including Favored Group

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Item Type	2	6.542	3.271	18693.516	0.000	0.000
DIF Magnitude	3	0.608	0.203	1157.609	0.000	0.011
Number of Clusters	1	2.757	2.757	15753.442	0.000	**0.134**
R:F Ratio	1	0.251	0.251	1433.279	0.000	0.012
Impact	1	0.046	0.046	260.100	0.000	0.002
ICC Magnitude	1	0.003	0.003	17.105	0.000	0.000
Group Favored	2	1.999	0.999	5710.507	0.000	0.000
Multilevel DIF Framework	2	2.362	1.181	6748.107	0.000	**0.115**
Item Type x DIF Magnitude	5	1.927	0.385	2202.283	0.000	0.000
Item Type x Number of Clusters	2	0.066	0.033	189.222	0.000	0.000
Item Type x R:F Ratio	2	0.007	0.003	19.949	0.000	0.000
Item Type x Impact	2	0.002	0.001	6.499	0.002	0.000
Item Type x ICC Magnitude	2	0.018	0.009	51.585	0.000	0.000
Item Type x Group Favored	1	0.013	0.013	73.915	0.000	0.000
Item Type x Multilevel DIF Framework	1	1.412	1.412	8069.888	0.000	0.000
DIF Magnitude x Number of Clusters	3	0.003	0.001	5.767	0.001	0.000
DIF Magnitude x R:F Ratio	3	0.002	0.001	2.977	0.030	0.000
DIF Magnitude x Impact	3	0.001	0.000	2.036	0.107	0.000
DIF Magnitude x ICC Magnitude	3	0.001	0.000	1.098	0.349	0.000
DIF Magnitude x Group Favored	2	0.001	0.001	3.766	0.023	0.000
DIF Magnitude x Multilevel DIF Framework	6	0.133	0.022	126.819	0.000	0.002
Number of Clusters x R:F Ratio	1	0.013	0.013	72.350	0.000	0.001
Number of Clusters x Impact	1	0.000	0.000	2.632	0.105	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Number of Clusters x ICC Magnitude	1	0.003	0.003	16.950	0.000	0.000
Number of Clusters x Group Favored	2	0.013	0.006	36.626	0.000	0.000
Number of Clusters x Multilevel DIF Framework	2	0.199	0.100	569.968	0.000	**0.010**
R:F Ratio x Impact	1	0.000	0.000	1.222	0.269	0.000
R:F Ratio x ICC Magnitude	1	0.000	0.000	1.408	0.235	0.000
R:F Ratio x Group Favored	2	0.002	0.001	5.498	0.004	0.000
R:F Ratio x Multilevel DIF Framework	2	0.013	0.006	36.614	0.000	0.001
Impact x ICC Magnitude	1	0.008	0.008	44.919	0.000	0.000
Impact x Group Favored	2	0.001	0.000	1.558	0.211	0.000
Impact x Multilevel DIF Framework	2	0.048	0.024	136.114	0.000	0.002
ICC Magnitude x Group Favored	2	0.001	0.001	3.466	0.031	0.000
ICC Magnitude x Multilevel DIF Framework	2	0.005	0.003	15.444	0.000	0.000
Group Favored x Multilevel DIF Framework	2	0.519	0.260	1483.052	0.000	0.000
Item Type x DIF Magnitude x Number of Clusters	5	0.002	0.000	2.573	0.025	0.000
Item Type x DIF Magnitude x R:F Ratio	5	0.000	0.000	0.092	0.993	0.000
Item Type x DIF Magnitude x Impact	5	0.000	0.000	0.442	0.819	0.000
Item Type x DIF Magnitude x ICC Magnitude	5	0.001	0.000	1.009	0.410	0.000
Item Type x DIF Magnitude x Group Favored	2	0.000	0.000	1.386	0.250	0.000
Item Type x DIF Magnitude x Multilevel DIF Framework	2	0.260	0.130	744.214	0.000	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Item Type x Number of Clusters x R:F Ratio	2	0.000	0.000	1.313	0.269	0.000
Item Type x Number of Clusters x Impact	2	0.000	0.000	0.401	0.670	0.000
Item Type x Number of Clusters x ICC Magnitude	2	0.000	0.000	0.837	0.433	0.000
Item Type x Number of Clusters x Group Favored	1	0.000	0.000	0.074	0.786	0.000
Item Type x Number of Clusters x Multilevel DIF Framework	1	0.014	0.014	78.247	0.000	0.000
Item Type x R:F Ratio x Impact	2	0.000	0.000	0.095	0.910	0.000
Item Type x R:F Ratio x ICC Magnitude	2	0.000	0.000	0.963	0.382	0.000
Item Type x R:F Ratio x Group Favored	1	0.001	0.001	4.999	0.025	0.000
Item Type x R:F Ratio x Multilevel DIF Framework	1	0.003	0.003	19.523	0.000	0.000
Item Type x Impact x ICC Magnitude	2	0.003	0.002	9.840	0.000	0.000
Item Type x Impact x Group Favored	1	0.008	0.008	47.916	0.000	0.000
Item Type x Impact x Multilevel DIF Framework	1	0.000	0.000	0.334	0.563	0.000
Item Type x ICC Magnitude x Group Favored	1	0.000	0.000	0.095	0.758	0.000
Item Type x ICC Magnitude x Multilevel DIF Framework	1	0.006	0.006	35.321	0.000	0.000
DIF Magnitude x Number of Clusters x R:F Ratio	3	0.000	0.000	0.871	0.455	0.000
DIF Magnitude x Number of Clusters x Impact	3	0.000	0.000	0.627	0.597	0.000
DIF Magnitude x Number of Clusters x ICC Magnitude	3	0.001	0.000	1.274	0.281	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
DIF Magnitude x Number of Clusters x Group Favored	2	0.000	0.000	0.096	0.908	0.000
DIF Magnitude x Number of Clusters x Multilevel DIF Framework	6	0.001	0.000	0.548	0.772	0.000
DIF Magnitude x R:F Ratio x Impact	3	0.000	0.000	0.552	0.647	0.000
DIF Magnitude x R:F Ratio x ICC Magnitude	3	0.000	0.000	0.319	0.811	0.000
DIF Magnitude x R:F Ratio x Group Favored	2	0.000	0.000	0.307	0.735	0.000
DIF Magnitude x R:F Ratio x Multilevel DIF Framework	6	0.001	0.000	0.942	0.463	0.000
DIF Magnitude x Impact x ICC Magnitude	3	0.000	0.000	0.007	0.999	0.000
DIF Magnitude x Impact x Group Favored	2	0.001	0.000	2.753	0.064	0.000
DIF Magnitude x Impact x Multilevel DIF Framework	6	0.001	0.000	0.964	0.448	0.000
DIF Magnitude x ICC Magnitude x Group Favored	2	0.000	0.000	0.051	0.950	0.000
DIF Magnitude x ICC Magnitude x Multilevel DIF Framework	6	0.000	0.000	0.336	0.918	0.000
DIF Magnitude x Group Favored x Multilevel DIF Framework	2	0.001	0.000	1.508	0.221	0.000
Number of Clusters x R:F Ratio x Impact	1	0.000	0.000	0.056	0.814	0.000
Number of Clusters x R:F Ratio x ICC Magnitude	1	0.000	0.000	1.014	0.314	0.000
Number of Clusters x R:F Ratio x Group Favored	2	0.000	0.000	0.980	0.375	0.000
Number of Clusters x R:F Ratio x Multilevel DIF Framework	2	0.000	0.000	0.502	0.605	0.000
Number of Clusters x Impact x ICC Magnitude	1	0.000	0.000	0.237	0.626	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Number of Clusters x Impact x Group Favored	2	0.000	0.000	0.046	0.955	0.000
Number of Clusters x Impact x Multilevel DIF Framework	2	0.003	0.002	8.575	0.000	0.000
Number of Clusters x ICC Magnitude x Group Favored	2	0.000	0.000	0.900	0.407	0.000
Number of Clusters x ICC Magnitude x Multilevel DIF Framework	2	0.002	0.001	6.630	0.001	0.000
Number of Clusters x Group Favored x Multilevel DIF Framework	2	0.003	0.002	9.135	0.000	0.000
R:F Ratio x Impact x ICC Magnitude	1	0.000	0.000	0.321	0.571	0.000
R:F Ratio x Impact x Group Favored	2	0.000	0.000	0.049	0.952	0.000
R:F Ratio x Impact x Multilevel DIF Framework	2	0.000	0.000	0.959	0.383	0.000
R:F Ratio x ICC Magnitude x Group Favored	2	0.000	0.000	0.022	0.978	0.000
R:F Ratio x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.495	0.610	0.000
R:F Ratio x Group Favored x Multilevel DIF Framework	2	0.001	0.001	4.011	0.018	0.000
Impact x ICC Magnitude x Group Favored	2	0.001	0.000	2.224	0.108	0.000
Impact x ICC Magnitude x Multilevel DIF Framework	2	0.015	0.007	42.819	0.000	0.000
Impact x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.082	0.921	0.001

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
ICC Magnitude x Group Favored x Multilevel DIF Framework	2	0.001	0.001	4.241	0.014	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio	5	0.000	0.000	0.232	0.949	0.000
Item Type x DIF Magnitude x Number of Clusters x Impact	5	0.000	0.000	0.142	0.982	0.000
Item Type x DIF Magnitude x Number of Clusters x ICC Magnitude	5	0.000	0.000	0.360	0.876	0.000
Item Type x DIF Magnitude x Number of Clusters x Group Favored	2	0.000	0.000	0.230	0.794	0.000
Item Type x DIF Magnitude x Number of Clusters x Multilevel DIF Framework	2	0.000	0.000	1.122	0.326	0.000
Item Type x DIF Magnitude x R:F Ratio x Impact	5	0.000	0.000	0.250	0.940	0.000
Item Type x DIF Magnitude x R:F Ratio x ICC Magnitude	5	0.000	0.000	0.015	1.000	0.000
Item Type x DIF Magnitude x R:F Ratio x Group Favored	2	0.000	0.000	0.011	0.989	0.000
Item Type x DIF Magnitude x R:F Ratio x Multilevel DIF Framework	2	0.000	0.000	0.916	0.400	0.000
Item Type x DIF Magnitude x Impact x ICC Magnitude	5	0.000	0.000	0.140	0.983	0.000
Item Type x DIF Magnitude x Impact x Group Favored	2	0.000	0.000	0.952	0.386	0.000
Item Type x DIF Magnitude x Impact x	2	0.000	0.000	1.053	0.349	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Multilevel DIF Framework						
Item Type x DIF Magnitude x ICC Magnitude x Group Favored	2	0.000	0.000	0.119	0.888	0.000
Item Type x DIF Magnitude x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.481	0.618	0.000
Item Type x Number of Clusters x R:F Ratio x Impact	2	0.000	0.000	0.563	0.570	0.000
Item Type x Number of Clusters x R:F Ratio x ICC Magnitude	2	0.000	0.000	0.056	0.946	0.000
Item Type x Number of Clusters x R:F Ratio x Group Favored	1	0.000	0.000	0.845	0.358	0.000
Item Type x Number of Clusters x R:F Ratio x Multilevel DIF Framework	1	0.001	0.001	2.933	0.087	0.000
Item Type x Number of Clusters x Impact x ICC Magnitude	2	0.000	0.000	0.952	0.386	0.000
Item Type x Number of Clusters x Impact x Group Favored	1	0.000	0.000	0.886	0.347	0.000
Item Type x Number of Clusters x Impact x Multilevel DIF Framework	1	0.000	0.000	0.534	0.465	0.000
Item Type x Number of Clusters x ICC Magnitude x Group Favored	1	0.000	0.000	0.695	0.405	0.000
Item Type x Number of Clusters x ICC Magnitude x Multilevel DIF Framework	1	0.000	0.000	0.727	0.394	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Item Type x R:F Ratio x Impact x ICC Magnitude	2	0.000	0.000	0.005	0.995	0.000
Item Type x R:F Ratio x Impact x Group Favored	1	0.000	0.000	0.381	0.537	0.000
Item Type x R:F Ratio x Impact x Multilevel DIF Framework	1	0.000	0.000	0.081	0.775	0.000
Item Type x R:F Ratio x ICC Magnitude x Group Favored	1	0.000	0.000	0.141	0.708	0.000
Item Type x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	1	0.000	0.000	0.245	0.620	0.000
Item Type x Impact x ICC Magnitude x Group Favored	1	0.000	0.000	0.096	0.757	0.000
Item Type x Impact x ICC Magnitude x Multilevel DIF Framework	1	0.006	0.006	36.438	0.000	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Impact	3	0.000	0.000	0.184	0.907	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x ICC Magnitude	3	0.000	0.000	0.724	0.538	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Group Favored	2	0.000	0.000	0.336	0.715	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Multilevel DIF Framework	6	0.000	0.000	0.337	0.918	0.000
DIF Magnitude x Number of Clusters x Impact x ICC Magnitude	3	0.000	0.000	0.525	0.665	0.000
DIF Magnitude x Number of Clusters x Impact x Group Favored	2	0.000	0.000	0.547	0.579	0.000
DIF Magnitude x Number of Clusters x Impact x	6	0.001	0.000	0.499	0.809	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Multilevel DIF Framework						
DIF Magnitude x Number of Clusters x ICC Magnitude x Group Favored	2	0.000	0.000	0.335	0.715	0.000
DIF Magnitude x Number of Clusters x ICC Magnitude x Multilevel DIF Framework	6	0.000	0.000	0.109	0.995	0.000
DIF Magnitude x Number of Clusters x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.216	0.806	0.000
DIF Magnitude x R:F Ratio x Impact x ICC Magnitude	3	0.001	0.000	1.366	0.251	0.000
DIF Magnitude x R:F Ratio x Impact x Group Favored	2	0.000	0.000	0.160	0.852	0.000
DIF Magnitude x R:F Ratio x Impact x Multilevel DIF Framework	6	0.001	0.000	0.952	0.456	0.000
DIF Magnitude x R:F Ratio x ICC Magnitude x Group Favored	2	0.000	0.000	0.437	0.646	0.000
DIF Magnitude x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	6	0.001	0.000	0.736	0.620	0.000
DIF Magnitude x R:F Ratio x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.120	0.887	0.000
DIF Magnitude x Impact x ICC Magnitude x Group Favored	2	0.000	0.000	0.158	0.853	0.000
DIF Magnitude x Impact x ICC Magnitude x Multilevel DIF Framework	6	0.000	0.000	0.218	0.971	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
DIF Magnitude x Impact x Group Favored x Multilevel DIF Framework	2	0.001	0.000	1.973	0.139	0.000
DIF Magnitude x ICC Magnitude x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.282	0.755	0.000
Number of Clusters x R:F Ratio x Impact x ICC Magnitude	1	0.000	0.000	0.015	0.903	0.000
Number of Clusters x R:F Ratio x Impact x Group Favored	2	0.000	0.000	0.273	0.761	0.000
Number of Clusters x R:F Ratio x Impact x Multilevel DIF Framework	2	0.000	0.000	0.344	0.709	0.000
Number of Clusters x R:F Ratio x ICC Magnitude x Group Favored	2	0.001	0.000	1.560	0.210	0.000
Number of Clusters x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.207	0.813	0.000
Number of Clusters x R:F Ratio x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.423	0.655	0.000
Number of Clusters x Impact x ICC Magnitude x Group Favored	2	0.000	0.000	0.212	0.809	0.000
Number of Clusters x Impact x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	1.081	0.339	0.000
Number of Clusters x Impact x Group Favored x Multilevel DIF Framework	2	0.000	0.000	1.132	0.322	0.000
Number of Clusters x ICC Magnitude x Group	2	0.000	0.000	0.129	0.879	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Favored x Multilevel DIF Framework						
R:F Ratio x Impact x ICC Magnitude x Group Favored	2	0.000	0.000	0.444	0.641	0.000
R:F Ratio x Impact x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.394	0.674	0.000
R:F Ratio x Impact x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.442	0.643	0.000
R:F Ratio x ICC Magnitude x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.020	0.980	0.000
Impact x ICC Magnitude x Group Favored x Multilevel DIF Framework	2	0.001	0.000	2.791	0.061	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x Impact	5	0.000	0.000	0.113	0.989	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x ICC Magnitude	5	0.000	0.000	0.133	0.985	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x Group Favored	2	0.000	0.000	0.165	0.848	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x Multilevel DIF Framework	2	0.000	0.000	0.095	0.909	0.000
Item Type x DIF Magnitude x Number of Clusters x Impact x ICC Magnitude	5	0.000	0.000	0.300	0.913	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Item Type x DIF Magnitude x Number of Clusters x Impact x Group Favored	2	0.000	0.000	0.022	0.978	0.000
Item Type x DIF Magnitude x Number of Clusters x Impact x Multilevel DIF Framework	2	0.000	0.000	0.152	0.859	0.000
Item Type x DIF Magnitude x Number of Clusters x ICC Magnitude x Group Favored	2	0.000	0.000	0.063	0.939	0.000
Item Type x DIF Magnitude x Number of Clusters x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.794	0.452	0.000
Item Type x DIF Magnitude x R:F Ratio x Impact x ICC Magnitude	5	0.000	0.000	0.139	0.983	0.000
Item Type x DIF Magnitude x R:F Ratio x Impact x Group Favored	2	0.000	0.000	0.117	0.889	0.000
Item Type x DIF Magnitude x R:F Ratio x Impact x Multilevel DIF Framework	2	0.000	0.000	0.297	0.743	0.000
Item Type x DIF Magnitude x R:F Ratio x ICC Magnitude x Group Favored	2	0.000	0.000	0.072	0.931	0.000
Item Type x DIF Magnitude x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.130	0.878	0.000
Item Type x DIF Magnitude x Impact x ICC Magnitude x Group Favored	2	0.000	0.000	0.166	0.847	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Item Type x DIF Magnitude x Impact x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.143	0.866	0.000
Item Type x Number of Clusters x R:F Ratio x Impact x ICC Magnitude	2	0.000	0.000	0.084	0.920	0.000
Item Type x Number of Clusters x R:F Ratio x Impact x Group Favored	1	0.000	0.000	0.000	0.994	0.000
Item Type x Number of Clusters x R:F Ratio x Impact x Multilevel DIF Framework	1	0.000	0.000	0.432	0.511	0.000
Item Type x Number of Clusters x R:F Ratio x ICC Magnitude x Group Favored	1	0.000	0.000	0.213	0.644	0.000
Item Type x Number of Clusters x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	1	0.000	0.000	0.123	0.726	0.000
Item Type x Number of Clusters x Impact x ICC Magnitude x Group Favored	1	0.000	0.000	0.380	0.537	0.000
Item Type x Number of Clusters x Impact x ICC Magnitude x Multilevel DIF Framework	1	0.000	0.000	1.780	0.182	0.000
Item Type x R:F Ratio x Impact x ICC Magnitude x Group Favored	1	0.000	0.000	1.064	0.302	0.000
Item Type x R:F Ratio x Impact x ICC Magnitude x Multilevel DIF Framework	1	0.000	0.000	0.036	0.849	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Impact x ICC Magnitude	3	0.000	0.000	0.178	0.911	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
DIF Magnitude x Number of Clusters x R:F Ratio x Impact x Group Favored	2	0.000	0.000	0.145	0.865	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Impact x Multilevel DIF Framework	6	0.000	0.000	0.195	0.978	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x ICC Magnitude x Group Favored	2	0.000	0.000	0.009	0.991	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	6	0.001	0.000	0.524	0.790	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.082	0.921	0.000
DIF Magnitude x Number of Clusters x Impact x ICC Magnitude x Group Favored	2	0.000	0.000	0.098	0.906	0.000
DIF Magnitude x Number of Clusters x Impact x ICC Magnitude x Multilevel DIF Framework	6	0.001	0.000	0.808	0.563	0.000
DIF Magnitude x Number of Clusters x Impact x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.119	0.888	0.000
DIF Magnitude x Number of Clusters x ICC Magnitude x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.663	0.515	0.000
DIF Magnitude x R:F Ratio x Impact x ICC	2	0.000	0.000	1.013	0.363	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Magnitude x Group Favored						
DIF Magnitude x R:F Ratio x Impact x ICC Magnitude x Multilevel DIF Framework	6	0.001	0.000	0.701	0.648	0.000
DIF Magnitude x R:F Ratio x Impact x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.076	0.927	0.000
DIF Magnitude x R:F Ratio x ICC Magnitude x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.111	0.895	0.000
DIF Magnitude x Impact x ICC Magnitude x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.247	0.781	0.000
Number of Clusters x R:F Ratio x Impact x ICC Magnitude x Group Favored	2	0.000	0.000	0.954	0.385	0.000
Number of Clusters x R:F Ratio x Impact x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.013	0.987	0.000
Number of Clusters x R:F Ratio x Impact x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.497	0.608	0.000
Number of Clusters x R:F Ratio x ICC Magnitude x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.415	0.660	0.000
Number of Clusters x Impact x ICC Magnitude x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.044	0.957	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
R:F Ratio x Impact x ICC Magnitude x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.039	0.962	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x Impact x ICC Magnitude	5	0.000	0.000	0.411	0.841	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x Impact x Group Favored	2	0.000	0.000	0.313	0.731	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x Impact x Multilevel DIF Framework	2	0.000	0.000	0.591	0.554	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x ICC Magnitude x Group Favored	2	0.000	0.000	0.088	0.916	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.025	0.975	0.000
Item Type x DIF Magnitude x Number of Clusters x Impact x ICC Magnitude x Group Favored	2	0.000	0.000	0.405	0.667	0.000
Item Type x DIF Magnitude x Number of Clusters x Impact x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.638	0.529	0.000
Item Type x DIF Magnitude x R:F Ratio x Impact x ICC Magnitude x Group Favored	2	0.000	0.000	0.341	0.711	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Item Type x DIF Magnitude x R:F Ratio x Impact x ICC Magnitude x Multilevel DIF Framework	2	0.001	0.000	1.780	0.169	0.000
Item Type x Number of Clusters x R:F Ratio x Impact x ICC Magnitude x Group Favored	1	0.000	0.000	0.135	0.713	0.000
Item Type x Number of Clusters x R:F Ratio x Impact x ICC Magnitude x Multilevel DIF Framework	1	0.000	0.000	0.139	0.710	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Impact x ICC Magnitude x Group Favored	2	0.000	0.000	0.060	0.942	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Impact x ICC Magnitude x Multilevel DIF Framework	6	0.000	0.000	0.276	0.949	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Impact x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.485	0.616	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x ICC Magnitude x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.080	0.923	0.000
DIF Magnitude x Number of Clusters x Impact x ICC Magnitude x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.415	0.660	0.000
DIF Magnitude x R:F Ratio x Impact x ICC Magnitude x Group	2	0.000	0.000	0.215	0.807	0.000

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Favored x Multilevel DIF Framework						
Number of Clusters x R:F Ratio x Impact x ICC Magnitude x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.438	0.646	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x Impact x ICC Magnitude x Group Favored	2	0.000	0.000	0.032	0.968	0.000
Item Type x DIF Magnitude x Number of Clusters x R:F Ratio x Impact x ICC Magnitude x Multilevel DIF Framework	2	0.000	0.000	0.092	0.912	0.000
DIF Magnitude x Number of Clusters x R:F Ratio x Impact x ICC Magnitude x Group Favored x Multilevel DIF Framework	2	0.000	0.000	0.473	0.623	0.000
Residuals	6896	1.207	0.000			
Total	7423	30.844				

APPENDIX D

IMPACT RESULTS

Table D.1

Relative Bias Results for Impact

		High ICC				Low ICC			
		Equivalent Abilities		Impact		Equivalent Abilities		Impact	
		Even	Uneven	Even	Uneven	Even	Uneven	Even	Uneven
Null (δ=0.0)	100	0.201	0.228	-0.022	0.006	0.099	0.104	-0.049	-0.070
	300	0.117	0.130	-0.037	-0.057	0.057	0.075	-0.080	-0.084
Small (δ=0.2)	100	0.178	0.223	-0.045	-0.070	0.091	0.106	-0.070	-0.079
	300	0.120	0.110	-0.025	-0.039	0.050	0.070	-0.086	-0.095
Medium (δ=0.4)	100	0.185	0.214	-0.022	0.018	0.104	0.109	-0.089	-0.080
	300	0.101	0.117	-0.048	-0.007	0.057	0.065	-0.088	-0.090
Large (δ=0.6)	100	0.176	0.210	-0.028	0.005	0.096	0.115	-0.104	-0.092
	300	0.103	0.118	-0.039	-0.046	0.060	0.071	-0.107	-0.093

Table D.2

RMSE Results for Impact

		High ICC				Low ICC			
		Equivalent Abilities		Impact		Equivalent Abilities		Impact	
		Even	Uneven	Even	Uneven	Even	Uneven	Even	Uneven
Null (δ=0.0)	100	0.266	0.290	0.239	0.283	0.123	0.130	0.142	0.162
	300	0.140	0.159	0.155	0.202	0.070	0.089	0.111	0.124
Small (δ=0.2)	100	0.217	0.286	0.213	0.238	0.113	0.134	0.153	0.170
	300	0.149	0.133	0.148	0.156	0.065	0.087	0.112	0.123
Medium (δ=0.4)	100	0.238	0.270	0.211	0.265	0.129	0.134	0.155	0.159
	300	0.125	0.147	0.143	0.158	0.076	0.082	0.122	0.123
Large (δ=0.6)	100	0.227	0.255	0.199	0.264	0.116	0.146	0.154	0.178
	300	0.132	0.147	0.143	0.173	0.076	0.088	0.133	0.132

Table D.3

The ANOVA Results for Relative Bias

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
DIF Magnitude	3	0.001	0.000	2.696	0.218	0.002
Social-Unit Sample Size	1	0.024	0.024	139.829	0.001	**0.039**
R:F Ratio	1	0.002	0.002	10.187	0.050	0.003
Impact Presence	1	0.503	0.503	2889.061	0.000	**0.803**
ICC Magnitude	1	0.069	0.069	395.144	0.000	**0.110**
DIF Magnitude x Social-Unit Sample Size	3	0.001	0.000	1.375	0.400	0.001
DIF Magnitude x R:F Ratio	3	0.001	0.000	1.504	0.373	0.001
DIF Magnitude x Impact Presence	3	0.000	0.000	0.790	0.575	0.001
DIF Magnitude x ICC Magnitude	3	0.000	0.000	0.801	0.570	0.001
Social-Unit Sample Size x R:F Ratio	1	0.000	0.000	2.212	0.234	0.001
Social-Unit Sample Size x Impact Presence	1	0.010	0.010	55.525	0.005	**0.015**
Social-Unit Sample Size x ICC Magnitude	1	0.003	0.003	16.462	0.027	0.005
R:F Ratio x Impact Presence	1	0.001	0.001	3.696	0.150	0.001
R:F Ratio x ICC Magnitude	1	0.000	0.000	2.131	0.240	0.001
Impact Presence x ICC Magnitude	1	0.001	0.001	8.185	0.065	0.002
DIF Magnitude x Social-Unit Sample Size x R:F Ratio	3	0.000	0.000	0.142	0.928	0.000
DIF Magnitude x Social-Unit Sample Size x Impact Presence	3	0.001	0.000	1.202	0.442	0.001
DIF Magnitude x Social-Unit Sample Size x ICC Magnitude	3	0.001	0.000	1.505	0.373	0.001

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
DIF Magnitude x R:F Ratio x Impact Presence	3	0.001	0.000	1.626	0.350	0.001
DIF Magnitude x R:F Ratio x ICC Magnitude	3	0.001	0.000	1.042	0.487	0.001
DIF Magnitude x Impact Presence x ICC Magnitude	3	0.002	0.001	3.879	0.147	0.003
Social-Unit Sample Size x R:F Ratio x Impact Presence	1	0.000	0.000	0.028	0.878	0.000
Social-Unit Sample Size x R:F Ratio x ICC Magnitude	1	0.001	0.001	3.577	0.155	0.001
Social-Unit Sample Size x Impact Presence x ICC Magnitude	1	0.002	0.002	9.758	0.052	0.003
R:F Ratio x Impact Presence x ICC Magnitude	1	0.000	0.000	0.029	0.875	0.000
DIF Magnitude x Social-Unit Sample Size x R:F Ratio x Impact Presence	3	0.000	0.000	0.558	0.678	0.000
DIF Magnitude x Social-Unit Sample Size x R:F Ratio x ICC Magnitude	3	0.000	0.000	0.442	0.740	0.000
DIF Magnitude x Social-Unit Sample Size x Impact Presence x ICC Magnitude	3	0.001	0.000	1.585	0.357	0.001
DIF Magnitude x R:F Ratio x Impact Presence x ICC Magnitude	3	0.000	0.000	0.351	0.794	0.000
Social-Unit Sample Size x R:F Ratio x Impact Presence x ICC Magnitude	1	0.000	0.000	0.076	0.801	0.000
Residuals	3	0.001	0.000			
Total	63	0.626				

Table D.4

The ANOVA Results for Root Mean Square Error

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
DIF Magnitude	3	0.001	0.000	3.040	0.193	0.006
Social-Unit Sample Size	1	0.078	0.078	593.032	0.000	**0.362**
R:F Ratio	1	0.007	0.007	56.784	0.005	**0.035**
Impact Presence	1	0.006	0.006	42.729	0.007	**0.026**
ICC Magnitude	1	0.094	0.094	716.377	0.000	**0.438**
DIF Magnitude x Social-Unit Sample Size	3	0.000	0.000	0.496	0.710	0.001
DIF Magnitude x R:F Ratio	3	0.000	0.000	0.445	0.738	0.001
DIF Magnitude x Impact Presence	3	0.000	0.000	0.350	0.794	0.001
DIF Magnitude x ICC Magnitude	3	0.002	0.001	6.318	0.082	**0.012**
Social-Unit Sample Size x R:F Ratio	1	0.001	0.001	6.967	0.078	0.004
Social-Unit Sample Size x Impact Presence	1	0.002	0.002	17.540	0.025	**0.011**
Social-Unit Sample Size x ICC Magnitude	1	0.012	0.012	89.086	0.003	**0.054**
R:F Ratio x Impact Presence	1	0.000	0.000	0.493	0.533	0.000
R:F Ratio x ICC Magnitude	1	0.001	0.001	8.988	0.058	0.005
Impact Presence x ICC Magnitude	1	0.005	0.005	40.502	0.008	**0.025**
DIF Magnitude x Social-Unit Sample Size x R:F Ratio	3	0.000	0.000	1.168	0.451	0.002
DIF Magnitude x Social-Unit Sample Size x Impact Presence	3	0.000	0.000	0.130	0.936	0.000
DIF Magnitude x Social-Unit Sample Size x ICC Magnitude	3	0.000	0.000	0.681	0.620	0.001

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
DIF Magnitude x R:F Ratio x Impact Presence	3	0.000	0.000	0.712	0.607	0.001
DIF Magnitude x R:F Ratio x ICC Magnitude	3	0.000	0.000	0.707	0.609	0.001
DIF Magnitude x Impact Presence x ICC Magnitude	3	0.000	0.000	0.586	0.664	0.001
Social-Unit Sample Size x R:F Ratio x Impact Presence	1	0.000	0.000	0.026	0.883	0.000
Social-Unit Sample Size x R:F Ratio x ICC Magnitude	1	0.000	0.000	2.879	0.188	0.002
Social-Unit Sample Size x Impact Presence x ICC Magnitude	1	0.001	0.001	4.030	0.138	0.002
R:F Ratio x Impact Presence x ICC Magnitude	1	0.000	0.000	1.917	0.260	0.001
DIF Magnitude x Social-Unit Sample Size x R:F Ratio x Impact Presence	3	0.000	0.000	0.844	0.554	0.002
DIF Magnitude x Social-Unit Sample Size x R:F Ratio x ICC Magnitude	3	0.000	0.000	0.931	0.523	0.002
DIF Magnitude x Social-Unit Sample Size x Impact Presence x ICC Magnitude	3	0.000	0.000	0.278	0.839	0.001
DIF Magnitude x R:F Ratio x Impact Presence x ICC Magnitude	3	0.000	0.000	0.475	0.722	0.001

Source	*df*	Sum of Squares	Mean Square	*F*	*p*	η^2
Social-Unit Sample Size x R:F Ratio x Impact Presence x ICC Magnitude	1	0.000	0.000	0.457	0.548	0.000
Residuals	3	0.000	0.000			
Total	63	0.216				