DETECTING AND EXPLAINING DIFFERENTIAL ITEM FUNCTIONING ON THE SOCIAL, ACADEMIC, AND EMOTIONAL BEHAVIOR RISK SCREENER

A Dissertation Presented to

the Faculty of the Graduate School

University of Missouri – Columbia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

Jared T. Izumi

University of Missouri – Columbia

Dr. Katie Eklund, Co-Advisor

Dr. Matthew K. Burns, Co-Advisor

July 2020

The undersigned, appointed by the dean of the Graduate School, have examined the dissertation entitled

DETECTING AND EXPLAINING DIFFERENTIAL ITEM FUNCTIONING ON THE SOCIAL, ACADEMIC, AND EMOTIONAL BEHAVIOR RISK SCREENER

Presented by Jared Izumi,

a candidate for the degree of doctor of philosophy, and hereby certify that, in their opinion, it is worthy of acceptance.

____

Professor Katie Eklund

____

Professor Matthew Burns

____

Professor Stephen Kilgus

____

Professor Wes Bonifay

____

Professor Erica Lembke

DEDICATION

This work is dedicated to my parents, without whom I would never have made it. I hope I can continue to make you as proud, as I am of you.

ACKNOWLEDGEMENTS

I am deeply grateful to co-advisors, Drs. Katie Eklund and Matt Burns, who gave me their time, expertise, and mentorship throughout my doctoral degree. I am also greatly appreciative of my other committee members that provided valuable insight and assistance throughout the project and graduate training. Your collective expertise in research and education have been and will continue to be extremely important as I pursue my own professional goals.

Next, I would like to thank the graduate students, preservice teachers, teachers, and administrators that directly made my dissertation project possible. I would not have been able to complete this project without your feedback and participation. I would like to acknowledge the Educational, School, and Counseling Psychology department for providing funding for this project.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Universal screening of social-emotional and behavioral (SEB) risk with teacher completed brief behavioral rating scales (BBRS) is one of the primary methods for identifying SEB risk in students. These measures should function similarly across races, ethnicities, and genders. However, there is limited research to support measurement invariance in universal screening for SEB risk. Therefore, the current study sought to expand upon the existing research on measurement invariance. The Emotional Behavior (EB) subscale of the Social, Academic, and Emotional Behavior Risk Screener-Teacher Rating Scale (SAEBRS-TRS) was examined. Measurement invariance was examined through differential item functioning (DIF) within item response theory (IRT). A unidimensional graded response model was fit to the data and indicated that effect sizes of DIF ranged from small to large for Black students compared to all non-Black students (Cohen's d = -0.11 to -0.87) and negligible to medium for White students compared to all non-White students (Cohen's d = -0.01 to 0.54). Effect sizes for Hispanic students and students with multiple races and ethnicities were small to negligible. Positively worded items and males had larger DIF effect sizes. Next, the Item Response Questionnaire (IRQ) was developed from information processes theory to compare the process teachers go through when completing questions on the EB subscale with the median absolute effect sizes. A micro-macro multilevel model was fit to the data and indicated that the IRQ was not a significant predictor of effect sizes. However, teachers' rank ordering of subjectivity of the EB subscale items were significantly negatively correlated with effect sizes. Limitations of the current study, implications for practice, and directions for future research are discussed.

CHAPTER I: INTRODUCTION

Background

Thirteen to 20% of all children live with a mental disorder at any given time, with prevalence rates expected to increase (Bor, Dean, Najman, & Hayatbakhsh, 2014; Perou et al., 2013). The highest rates of mental disorders include social-emotional (e.g., mood and anxiety disorders) and behavioral disorders (e.g., ADHD, oppositional defiant disorder, and conduct disorder; Perou et al., 2013). Social-emotional and behavioral (SEB) problems are associated with increased school difficulties including reduced academic achievement, and poor social and emotional functioning (Hinshaw, 1992; King, Lembke, & Reinke, 2015). Early identification and intervention offer potential solutions to help improve outcomes for children, as untreated problems may become resistant to change over time (Kratochwill, 2007). One way to identify these children early is through universal screening.

Universal screening for SEB problems is supported by current federal legislation. The Every Student Succeeds Act (ESSA; 2015) expands the availability of comprehensive services that are provided to students, which includes early identification. In addition, the reauthorization of the Individuals with Disability Education Act (IDEA, 2004) mandates that schools engage in early identification strategies through child find requirements that seek to identify children who may demonstrate evidence of barriers to learning.

Schools have used proactive and reactive methods to identify individuals with SEB risk (Dowdy, Doane, Eklund, & Dever, 2011; McIntosh, Campbell, Carter, & Zumbo, 2009). Reactive methods of SEB risk identification could include using existing student information to examine rates of SEB risk (e.g., office discipline referrals [ODRs], suspensions, expulsions, referrals for special education). Research has consistently noted that minority students are disproportionately represented in some reactive screening methods such as disciplinary referrals and referrals for special education services (American Psychological Association Zero Tolerance Task Force, 2008; National Research Council, 2002).

In contrast, schools can use proactive methods of identifying SEB risk (e.g., brief behavioral rating scales [BBRS] and systematic teacher nominations; Severson, Walker, Hope-Doolittle, Kratochwill, & Gresham, 2007). Proactive methods attempt to objectify the identification process by recognizing early risk factors associated with poor SEB functioning (Dowdy, Ritchey, & Kamphaus, 2010). Objective measures of behavioral functioning have been suggested as a means to reduce the disproportionate identification of at-risk students, across culturally and linguistically diverse student populations (Raines, Dever, Kamphaus, & Roach, 2012). However, BBRS of SEB functioning maintain some level of subjectivity as individuals are required to interpret each question by making a judgment about the behavior in question and its frequency, intensity, and topography.

Statement of the Problem

The subjectivity in BBRS can be particularly difficult in the United States as it is one of the most culturally and linguistically diverse countries in the world (Banks, 2015). Significant between group differences may be found because of the moderating effects of diversity (Cook, Volpe, & Livanis, 2010). A moderator in research is a variable that changes the direction and/or strength of the relationship between the independent variable and dependent variable (Baron & Kenny, 1986). For example, race/ethnicity, biological sex, age, grade, and latent trait level can moderate outcomes in screening research. On rating scales, the effects of moderation can be categorized into two types: (1) item impact, actual differences that occur as a function of the moderating variable; and (2) item bias, differences due to an underlying characteristic of the question or measure that occurs as a function of the moderating variable (Zumbo, 2007). Due to the heterogeneity of schools in the United States, it would be problematic to assume that BBRS that are used in universal screening function without moderating effects. Researchers have begun to examine moderating effects in BBRS used for universal screening of SEB risk (Dowdy, Dever, DiStefano, & Chin, 2011; Lambert, January, Cress, Epstein, & Cullinan, 2018; Schatschneider, Lane, Oakes, & Kalberg, 2014); however, additional research is needed to determine the effects of race/ethnicity and the interaction of race/ethnicity and biological sex on SEB risk identification through the use of BBRS. Findings from this study investigation will provide implications for BBRS that are used as universal screening tools for SEB problems. The results could also inform future SEB rating scale development that are used as universal screening tools.

Purpose of the Current Study

Research is needed to evaluate the moderating effects of diversity on SEB risk identification with BBRS that are used for universal screening (Cook et al., 2010). Current research on reactive methods for risk identification including disciplinary practices and teacher referrals for special education indicate that students from minority backgrounds are disproportionately identified. Many explanations have been provided regarding disproportionality, including subjective interpretations of behaviors and application of those interpretations on BBRS (Tourangeau & Rasinski, 1988; Townsend, 2000). It is important to consider how raters process items and provide responses on rating scales when evaluating if students from different groups are being treated equally as part of a universal screening process for SEB risk. Therefore, the purpose of the current study was to examine a BBRS that is used as a universal screener for SEB risk to:

1. Describe the frequencies of SEB risk by race/ethnicity and the interaction of biological sex and race/ethnicity
2. Evaluate if individuals are being treated similarly regardless of group membership (i.e., biological sex and race/ethnicity) by identifying items that display differential item functioning (DIF) on a teacher completed version of a BBRS of SEB functioning used for universal screening purposes
3. Identify trends in items that display DIF by group membership
4. Predict which items will display DIF on a SEB BBRS rating scale used for universal screening purposes by examining the subjectivity within each item.

Definitions of Key Terms

Universal Screening: Evaluation of all individuals within a given population (e.g., schools) for the purpose of identifying individuals at-risk and health of the system (Dowdy et al., 2015).

Social-emotional and behavior (SEB): SEB skills are a broad group of externalizing, internalizing, and adaptive competences that facilitate resilience and adaptation in the presence of stressors (Kamphaus, 2012)

Emotional behavior (EB): EB is one of the subscales of the SAEBRS-TRS refers to the ability to regulate emotion, adapt to changes, and respond to stressful events (Kilgus, Sims, von der Embse, & Taylor, 2015).

Information processing theory (IPT): IPT is a model that describes the process that individuals go through when completing rating scales based off attitudes (Tourangeau and Rasinski,1988).

Differential item functioning (DIF). DIF is a form of measurement invariance to identify if a measure if functioning equally across subgroups of individuals (Zumbo, 2007).

Assumptions

There are several assumptions in the current study. First, teachers completed rating of their students with the SAEBRS-TRS. These data were collected previous to the start of the current study. It was assumed that teachers rated their students to the best of their ability, and the data were collected and recorded accurately. Second, the current study used item response theory (IRT), which has different assumptions that those used in classical test theory (Reise et al., 2005). The assumption of invariance in IRT states that item properties (e.g., discrimination and threshold parameters) are not dependent on the particular characteristics of the calibration sample. The items were calibrated from a large representative sample of students, and it was assumed that the item properties would hold with similar samples. Therefore, an independent sample of teachers were recruited for study two from schools that were already using the SAEBRS as part of their school-based practice.

Delimitation

First, the teachers in study two were recruited through a single school district, and may not match the response of a more diverse sample. Second, the current study was limited to the Emotional Behavior (EB) subscale of SAEBRS. This limits the generalization of findings to other areas of SEB functioning. In addition, the SAEBRS was developed with a bifactor model, but the current study used a unidimensional model because only one subscale was examined. Multidimensional calibration of the items may change the effect sizes identified in this study.

CHAPTER II: LITERATURE REVIEW

In this chapter I review the literature surrounding the project's purpose. First, I discuss universal screening for social-emotional and behavioral risk and methods used within universal screening. Next, I describe the disproportionality in different methods and possible explanations. Lastly, I describe research on measurement invariance with different universal screening measures.

Prevention

Assessment and individual differences have a long history in the field of psychology starting with the use of intelligence, personality, and other mental health assessments (Benjamin, 2014). Psychological assessments have focused on evaluating intraindividual and interindividual differences. Clinicians attempt to uncover the nature and extent of intra- and inter-individual differences during individual assessment by using evaluation tools including records, interviews, observations, and standardized tests (American Education Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014).

In schools, individuals are typically being evaluated to determine if they meet a specified level of impairment, and therefore requires special education and related services (Merrell, Ervin, & Peacock, 2012). However, this is a narrow view of the role of assessment in schools and does not include the variety of assessment methods that school psychologists use as part of data-based decision-making (e.g., progress monitoring, program evaluation, needs assessment, and screening; Benson et al., 2019). The data that are gathered are often used to identify problems, determine the cause(s) of the problems, inform interventions, and evaluate the progress and recommend changes to the interventions across levels (e.g., student, classroom, and school) as part of a problem-solving process (Tilly, 2002).

Although the field of school psychology has broadened, practitioners have had difficulty transitioning from the role of gatekeeper (Merrell et al., 2012). Comprehensive evaluations are needed for certain individuals; however, using comprehensive one-on-one evaluations to prevent or mitigate future problems can be inefficient and costly (Chatterji, Caffray, Crowe, Freeman, & Jensen, 2004; Dowdy, Ritchey, & Kamphaus, 2010). In addition, the medical model of assessment may be one of many factors that has led to under-service of students across different groups (Carter et al., 2004; Walker et al., 2001).

A prevention-oriented model may resolve or mitigate early difficulties that otherwise would lead to significant impairment in the future (Carter, Briggs-Gowan, & Ornstein, 2004; Coie et al., 1993). Epidemiology is a core component of a prevention-oriented model because it describes the rate and distribution of diseases, and risk across a population (Herman, Riley-Tillman, & Reinke, 2012). A prevention-oriented model of surveillance is not new. For example, a prevention-oriented model of early identification and intervention through screening has been mandated as part of Medicaid since 1967 (i.e., Early and Periodic Screening, Diagnostic and Treatment; U.S. Department of Health and Human Services, Centers for Medicare & Medicaid Services, 2015). Multiple public health approaches have been suggested as a means by which to identify early signs and symptoms of risk paired with the delivery of evidence-based interventions. More recently, these efforts have been adopted to identify SEB risk in schools (Bruhn, Woods-Groves, & Huddle, 2014). In schools, universal screening is one part of a larger system of tiered assessment and intervention.

Prevention Frameworks

Surveillance data assess a broad and/or narrow range of needs across all individuals to monitor the health of that population and to inform data-based decision making (Dowdy et al., 2015). Surveillance practices are a core component of prevention (Herman et al., 2012), and will be referred to as universal screening in this manuscript. Schools that use universal screening typically implement the practice within a multi-tiered model of assessment and intervention. Universal screening fits well within a tiered model because it supports the use of evidence-based practices and maximizes resource allocation (Severson et al., 2007). Schools implement a variety of multitiered systems of supports (MTSS) such as response to intervention (RTI) and positive behavior interventions and supports (PBIS; Fuchs & Fuchs, 2006; Severson et al., 2007; Sugai & Horner, 2002). MTSS use three levels or tiers of assessment and intervention (Severson et al., 2007), and with each successive level, assessments and interventions become more intense, specific, and comprehensive (Fuchs & Fuchs, 2006; Lane et al., 2015; Sugai & Horner, 2002).

At the first tier or universal level, all individuals are provided with evidence-based practices (Mellard, Stern, & Woods, 2011). This may include such services as school-wide instruction of behavioral expectations, evidence-based teaching strategies, and character education (Lane et al., 2015). Universal screening is also used to determine the overall health of the system, and to identify individuals that are at-risk for problematic academic and/or social-emotional and behavior (SEB) outcomes (Dowdy et al., 2015). Approximately 75-80% of all individuals should have their needs met by the natural supports and services that are available to all students (Severson et al., 2007;

Sugai & Horner, 2007). The remaining 20-25% of individuals may be identified as at-risk and would require additional services at the second tier or targeted level (Fuchs & Fuchs, 2006). Individuals requiring targeted intervention would be provided additional evidence-based academic and/or SEB services such as small group interventions or other complementary services (Mellard et al., 2011).

Approximately 1-10% of individuals that are not successful at the second tier of intervention would then require additional tier three or intensive assessment and intervention support (Burns, Appleton, & Stehouwer, 2005; Fuchs & Fuchs, 2006; Severson et al., 2007). These individuals typically display the most problematic academic and/or SEB problems that are the most resistant to evidence-based intervention (Sugai, Sprague, Horner, & Walker, 2000). As such, the role of universal screening within MTSS is used to determine the overall health of the system (e.g., school or district) and to act as an impetus to intervention by identifying those children at-risk for poor academic, social-emotional, and behavioral outcomes.

Research has demonstrated the positive effects of MTSS (Burns, Appleton, & Stehouwer, 2005; Gage, Whitford, & Katsiyannis, 2018; Solomon, Klein, Hintze, Cressey, & Peller, 2012). For example, a meta-analysis of RTI indicated that both large-scale and researcher implemented RTI resulted in improved academic and systemic outcomes (e.g., referral and placement in special education, time in special education services, and the number of students retained; Burns et al., 2005). The researchers found unbiased estimates of effect were greater than 1.0 for both researcher and large-scale implemented. However, other another student found negligible to negative effects of RTI on academic outcomes (Balu et al., 2015). The researchers found one statistically significant negative effect size in reading for first grade students. Students that were assigned to either tier 2 or tier 3 intervention had lower reading score outcomes than students that did not receive intervention (effect size of -0.17). A single case design and a group-based experimental meta-analyses have been conducted on the effectiveness of PBIS (Gage et al., 2018; Solomon et al., 2012). Both studies found nonsignificant reductions in office discipline referrals, but suspensions were reduced for students (g = -0.86; Gage et al., 2018). Universal screening can be used within a greater prevention model to provide targeted and indicated intervention.

Universal Screening

Universal screening is a proactive method of evaluating the health and condition of a system, rather than an individual, by assessing the functioning of all of the individuals within that system (Bowers, 1974; Fuchs & Fuchs, 2006; Mellard et al., 2011). The focus of universal screening is to promote the health and wellbeing of a community by preventing or reducing the intensity, frequency, or duration of problems within that system (Centers for Disease Control Foundation, n.d.). Schools serve as one context where universal screening measures can be administered. Schools may be one of the best settings to conduct universal screening given the large number of children and youth that attend schools (Coei et al., 1993; Farmer et al., 2004). In fact, many schools are already engaged in some form of universal screening practice, including screening for hearing, vision, and academic achievement (e.g., Green et al., 2013, Kemper, Fant, Bruckman, & Clark, 2004; Lane et al., 2015; Snyder, de Brey, & Dillow, 2019).

Within the broader MTSS framework, academic screening has far outpaced SEB screening. Mellard and colleagues (2009) surveyed 41 schools on their academic screening practices and found that 90% of schools were using three or more academic universal screening tools. In contrast, a separate study found that one in eight schools conducted universal screening for SEB problems (Bruhn et al., 2014). Many schools have implemented universal prevention efforts for academic difficulties, but continue to rely on reactionary disciplinary practices for SEB problems (Lane et al., 2015). Although efforts to conduct universal screening for SEB problems have increased from 2% of schools in 2005 to 12.6% in 2014 (Bruhn et al., 2014; McIntosh & Romer, 2005), several factors have been associated with the slow adoption of universal screening for SEB disorders including, the high cost of measures, lack of tools and procedures for universal screening, lack of awareness of positive outcomes associated with screening, stigma associated with identifying SEB problems related to psychopathology in children, lack of service providers for identified individuals, and system-level problems (Arora et al., 2016; Carter et al., 2004; Chatterji et al., 2004; DiStephano & Kamphaus, 2007; Harrison, Vannest, & Reynolds, 2013; Hartman et al., 2017).

Universal screening for SEB risk has been used to improve school outcomes (e.g., Cook et al., 2015; McIntosh, Chard, Boland, & Horner, 2006) while meeting the needs of more students and meeting their needs with less cost to society (Chatterji et al., 2004). For example, Chatterji and colleagues (2004) found that the total cost of screening and treating students in schools was less than the cost to society over a three-year period, and ranged from 8% to 24% lower cost at school than society. In another study, McIntosh and colleagues (2006) found that universal screening for academics and behavior along with interventions had greater reading proficiency and less office discipline referrals than national norms. Similarly, research found that providing PBIS, universal social-emotional learning, or a combined approach resulted in improved externalizing and internalizing problem behaviors compared to a business as usual group (Cook et al., 2015). For example, in the combined group, there was a large effect on decreasing externalizing and internalizing problem behaviors (i.e., externalizing Cohen's $d = 1.12$ and internalizing Cohen's $d = 0.74$).

Researchers have continued to make significant strides in the area of SEB screening practices, which includes the development or update of brief behavioral rating BBRS used in universal screening for SEB problems (e.g., the Social, Academic, and Emotional Behavior Risk Screener [SAEBRS], Student Risk Screening Scale for Internalizing and Externalizing [SSRS-IE]; Behavioral and Emotional Screening System [BESS]). BBRS as well as other forms of universal screening (e.g., office discipline referrals [ODRs]) can be used to predict end of year outcomes (Eklund, Kilgus, von der Embse, Beardmore, & Tanner, 2017; McIntosh, Frank, & Spaulding, 2010). For example, the SAEBRS subscales and Total Behavior from the beginning of the school year were correlated with end of year reading scores ranging from .16 o .40. In another study, McIntosh and colleagues (2010) found that by October, students with two or more ODRs were 176 times more likely to have gang affiliations displayed (OR = 175.81).

Methods of Identifying SEB Risk

Several methods have been used to identify individuals with SEB concerns (e.g., records, interviews, surveys, direct assessment, and observations) as well as the various informants available to provide information about students (e.g., self-report, parents, and teachers; Carter et al., 2004). Currently, three methods have been used to identify individuals that are at-risk for or are currently displaying significant distress related to

SEB problems. This includes office discipline referrals (ODRs), teacher nominations, and rating scales (Dowdy et al., 2010). An outline of the strengths and limitations of each of these methods is important when attempting to understand how these different methods may influence SEB risk identification.

Office Discipline Referrals (ODRs)

ODRs are one method schools use for collecting behavior data. Sugai and colleagues (2000) define an ODR as:

> An event in which (a) a student engages in a behavior that violated a rule or social norm in the school, (b) the problem behavior was observed or identified by a member of the school staff, and (c) the event resulted in a consequence delivered by administrative staff who produced a permanent (written) product defining the whole event (p. 96).

Data suggests that most schools already collect ODR information (Predy et al., 2014; Sugai et al., 2000), making it a relatively easy metric to assess discipline events in the school setting. However, the subjective nature of ODRs significantly limits their application (Girvan, Gion, McIntosh, & Smolkowski, 2017; Miller et al., 2015; Sugai et al., 2000). That is, one staff member within a school may determine a behavior meets the level of which to administer an ODR, whereas the same behavior may evoke a different response from a second individual. Another limitation of using ODRs are the reactive nature in which the data are used (McIntosh, Frank, & Spaulding, 2010). That is, students have to receive an ODR and typically multiple infractions before the student can be identified as a student that may require additional supports or services. A significant problem with this process is that an intervention based off ODR data is often implemented too late and a problematic relationship between the student and teacher has been established (McIntosh et al., 2010).

Research has found that ODRs are more sensitive to behavioral differences in students compared to other data collection methods using records (i.e., detention, suspension, and expulsion data; Sugai et al., 2000). However, other researchers have noted that ODRs are less sensitive than other methods of measurement (e.g., rating scales) for individuals with few ODRs (McIntosh et al., 2010; Predy, McIntosh, & Frank, 2014). ODRs have been used to identify individuals with externalizing problems, but may not capture the full spectrum of SEB problems (e.g., internalizing behaviors; Irvin et al., 2004; Martella et al., 2010; McIntosh et al., 2009; Miller et al., 2015; Predy et al., 2014; Severson et al., 2007).

Teacher Nomination

Teacher identification of student concerns and referral for support services have been the primary method by which students are identified for SEB problems (Dowdy, Doane, Eklund, & Dever, 2011; Gerber & Semmel, 1984; National Research Council, 2002). In this method, a teacher may identify some type of SEB concern in the classroom and would refer a student for additional help or support to a team or professional in the school tasked with providing student-focused interventions.

One tool that uses a systematic method to nominate students with SEB problems is the Systematic Screener for Behavioral Disorders (SSBD; Walker & Severson, 2014). The SSBD uses a multiple gating method to identify students with externalizing and internalizing behavior problems. At the first stage, teachers rank order students in their classroom to identify students demonstrating the highest levels of externalizing and internalizing behaviors. The teacher then selects three students with the highest externalizing problems and three students with the highest internalizing problems. These students are evaluated using a behavioral rating tool followed by direct observations. The SSBD has been shown to demonstrate adequate reliability with externalizing (α = .75), adaptive (α = .83), and maladaptive behaviors (α = .90) and convergent validity with the Teacher Rating Form and Social Skills Rating System (correlations between .47 and .67 on similar scales, with correlations below .38 on dissimilar scales; (Caldarella, Young, Richardson, Young, & Young, 2008; Richardson, Caldarella, Young, Young & Young, 2009; Walker et al., 1990).

Research regarding SSBD has not compared individuals at Stage One that are provided further evaluation and those individuals that are eventually identified in later gates. Intensity, duration, frequency, and topography of behaviors will influence how teachers rank behaviors. In addition, teachers may be misidentifying the rankings specifically for internalizing behaviors because teachers demonstrate more difficulties identifying students with internalizing behavior problems (De Los Reyes et al., 2015; Herman et al., 2018; Lochman, 1995). For example, a comparison of students at rank three and four may indicate there are no differences in their behavioral functioning (e.g., ODRs or Stage Two teacher ratings). It may be that most false negatives (e.g., the SSDB does not identify the individual as at-risk when the individual is at-risk) are those individuals just outside of the top three rankings in either externalizing or internalizing problems. In addition, internal consistency was low for internalizing problems (Cronbach's α = .57; Caldarella et al., 2008). The alpha level is well below reliability level of .70 for low-stakes decisions, like those done for screening purposes (Cortina, 1993; Jonnson & Svingby, 2007).

In a study using a less formalized methods of teacher nomination of students at-risk for SEB, researchers revealed no differences between students that were nominated by teachers for SEB risk and students that were not nominated for SEB risk on end-of-year ODR data and grades in reading (Dowdy et al., 2011). However, the study did not compare nominations for different areas of SEB problems (e.g., internalizing, externalizing, and school problems). As teachers tend to identify students who demonstrate more externalizing problems (e.g., aggressive, hyperactive, and disruptive behaviors) at higher rates than internalizing problems (e.g., depression and anxiety; Eklund & Dowdy, 2014; Lloyd et al., 1991; Pearcy, Clopton, & Pope, 1993), using a method such as the SSBD may demonstrate inherent flaws. Without a systematic approach to teacher nominations, a significant number of individuals may be under-identified or under-served (Dowdy et al., 2013; Eklund et al., 2009; Miller et al., 2015).

Rating Scales

A third method of SEB risk identification is the use of rating scales. Universal screening measures are one example of rating scales; these measures are designed to be brief assessments of student functioning that operationalize and structure respondents' perceptions on a broad range of SEB indicators (Dowdy et al., 2013; Eklund et al., 2017; Kamphaus et al., 2007). Brief behavioral rating scales (BBRS) are a more objective method of identifying SEB risk because each individual is evaluated on the same criteria; these measures often serve to identify additional students that are at-risk for SEB problems that other methods may miss (e.g., ODRs and nominations; Chatterji et al., 2004; Dowdy et al., 2013; Eklund et al., 2009; Eklund & Dowdy, 2014). BBRS of SEB problems have been found to identify individuals with externalizing problems, internalizing problems, adaptive problems, and school problems (Kilgus, Eklund, von der Embse, Taylor, & Sims, 2016; Kilgus, Taylor, & von der Embse, 2017; Stiffler & Dever, 2015). For example, Kilgus and colleagues (2016) found medium to large effects on academic outcomes when comparing students at-risk and not at-risk on the Social, Academic, and Emotional Behavior Risk Screener (SAEBRS) and sensitivity of >.90 and specificity of greater than .70 with other measures of SEB risk.

BBRS have also been found to be predictive of end of year behavioral data and academic data, beyond what can be explained with academic data (Eklund et al., 2017). In their study, Eklund and colleagues (2017) found that the individual scales of the SAEBRS accounted for 21% of the variability in reading curriculum-based measurement scores. BBRS screening data allow for schools to put in place early intervention services for students before SEB problems manifest into larger problems (Dowdy et al., 2010, 2013; Eklund et al., 2017). BBRS can be completed by different stakeholders, each demonstrating their own advantages and disadvantages.

Parents as informants. Parents routinely complete rating scales about their children. Parents can be a particularly important source of information when children are starting school (e.g., preschool and kindergarten) as they may serve as the most knowledgeable informant of what is typical/atypical for their child (Puura et al., 1998). From a pragmatic standpoint, teachers may not have had the time to learn about the student (Reynolds & Kamphaus, 2015) and self-reports would be impractical at this age (Levitt et al., 2007). Caution should be used as research demonstrates parents may underreport SEB problems due to stigma or cultural differences (Carter et al., 2004). For example, they may want to portray their child positively because they may perceive that poor SEB functioning of their child reflects their parenting ability. In addition, ratings by parents may not reflect SEB problems displayed in contexts other than those observed (e.g., behaviors outside of the home environment; Carter et al., 2004; Girio-Herrera et al., 2015), and their perspective of SEB functioning at school may not be as reliable or valid as teacher ratings (Girio-Herrera et al., 2015). Relatedly, research has consistently shown that parents are better able to identify their child's externalizing behaviors over that of their child's internalizing behaviors (De Los Reyes et al., 2015; Herman et al., 2018; Puura et al., 1998).

Teachers as informants. Teachers are one of the most common sources of information when collecting data on the SEB functioning of students in schools. Teacher ratings of SEB functioning demonstrate strong reliability and validity coefficients (Eklund et al., 2017; Kamphaus et al., 2007). For example, the SAEBRS had Cronbach's alphas of .94 for the Total Behavior scale with elementary and middle school students, and had subscale Cronbach's alpha ranging from .81 to .93 (Eklund et al., 2017), which is above the acceptable criterion of .80 for low-stakes decision making (Salvia, Ysseldyke, & Witmer, 2016).

Teacher ratings predict important behavioral and academic outcomes (Dowdy et al., 2013; Eklund et al., 2017; Kamphaus et al., 2007). Eklund et al. (2017) used a multilevel approach to evaluate a SEB screening tool with teachers and found little variability between teachers in how they rated the behavior of their students. The low variability may be due to reduced teacher bias and increased objectivity of rating scales (i.e., each teacher is rating their students' behavior similarly) or limited test sensitivity to actual differences in behavior (i.e., teachers rate their students similarly even though the students display varying intensity, frequency, or duration of behaviors; Eklund et al., 2017). However, teachers are in a unique role as they often have a normative comparison group in that they are able to evaluate a student's behavior against other students in the classroom and school, and therefore may be more sensitive to differences in SEB functioning (Puura et al., 1998). However, like parents, teachers are generally better at identifying student externalizing behavioral concerns than internalizing behaviors (De Los Reyes et al., 2015; Herman et al., 2018; Lochman, 1995).

Student self-report as informants. Students self-reports of SEB functioning can be an important source of information once students are mature enough to understand the constructs being measured (Levitt et al., 2007), with self-report measures developed for children in grades 3 and above (e.g., BASC-3 BESS; Kamphaus & Reynolds, 2015). Children may be better able to identify covert peer interaction behaviors (e.g., bullying between peers) or internal states (e.g., worry and sadness) better than parents or teachers (De Los Reyes et al., 2015). Student self-report data predict later behavioral and academic outcomes (Carroll et al., 2009; von der Embse, Kilgus, Iaccarino, & Levi-Nielsen, 2017). For example, von der Embse and colleagues (2017) found that the Total Behavior scale of the SAEBRS self-report version had adequate sensitivity and specificity and was highly correlated with another measure of behavioral functioning ($r > .50$). However, the Total Behavior scale had a Cronbach's alpha of .80, and all subscales had alpha scores between .63 and .68. The Total Behavior scale met the recommended

.80 criterion for reliability, but the subscales did not meet this criterion (Salvia et al., 2016).

Universal screening of SEB problems through rating scales may be more difficult to implement compared to teacher reports due to the extra protections afforded through the Protection of Pupil Rights Amendment (2002). These guidelines suggest that if screening procedures are voluntary and the student and/or their parents are allowed to opt out of the process, then self-reports can be a practical and psychometrically sound method of universal screening for SEB risk (Dever, Kamphaus, Dowdy, Raines, & DiStefano, 2013; Raines et al., 2012).

Synthesis

Prevention-oriented assessment in schools in the form of universal screening is a method that provides population level data on the overall health of the system on a broad and/or narrow range of needs. Universal screening is embedded at the first tier within a multi-tiered framework of assessment and intervention. Recently, this method has been used to identify SEB risk in students. Schools can use three main methods of collecting universal screening data on SEB risk: (1) through existing permanent records like ODRs, (2) systematics nominations by teachers, and (3) standardized BBRS. ODRs have been shown to be predictive of end of year outcomes; however, due to their reactive nature, interventions may be implemented too late when a negative relationship has been established between the teacher and student. In addition, this method may miss students that display internalizing behaviors. Teacher nominations have shown concurrent reliability and validity with rating scales. However, these methods rank students by perceptions of severity rather than classify students against a criterion. Research has found that BBRS identified additional students at-risk for SEB problems. BBRS can be completed by parents, teachers, and student self-reports. Parent reports are most beneficial at the beginning of a child's schooling (e.g., kindergarten), teacher reports are beneficial across a student's education, but can become more difficult to implement in secondary education when students have more than one teacher, and student self-reports are best when they are able to perceive and convey their own states of functioning (e.g., in third grade or higher). Student self-reports may provide the best information when examining internalizing behaviors, but may be more difficult to implement due to the additional protections provided to them.

Overall, the three methods can be used to identify SEB risk, but additional research is needed to understand the moderating effects of diversity in SEB identification (Cook et al., 2010). Specifically, researchers have found that ethnic and racial minorities are disproportionally identified as at-risk for SEB problems compared to their White counterparts (e.g., Bradshaw, Mitchell, O'Brennan, & Leaf, 2010b; Skiba et al., 2002, 2011; Smolkowski et al., 2016).

Disproportionality and Universal Screening

Disproportionality can be defined as the unequal distribution of individuals, through either overrepresentation or underrepresentation, measured on a construct or setting of interest (Raines, Dever, Kamphaus, & Roach, 2012; Raines, 2016). The unequal distribution is compared to the overall proportion of individuals in that population. In research and practice, disproportionality is typically defined by one group's risk in relation to another group's risk (e.g., risk ratios, odds ratios, and comparison index; Bottiani, Bradshaw, & Gregory, 2018). Risk ratios are defined as the relative risk of one group divided by the relative risk of a comparison group. For example, the risk ratio of Black students receiving an office discipline referral in relation to White students can be calculated by:

$$Relative\ Risk\ Ratio = \frac{\left(\frac{Number\ of\ Black\ students\ wiht\ ODRs}{Total\ number\ of\ Black\ students}\right)}{\left(\frac{Number\ of\ White\ students\ with\ ODRs}{Total\ number\ of\ White\ students}\right)} \tag{1}$$

A relative strength of the relative risk ratio is that it is easy to understand, in that it signifies if a group is represented on a variable at similar or different levels compared to another group (e.g., Hispanic students are twice as likely to be identified as having a Learning Disability compared to White students). However, a significant limitation of this metric is the reliance on a comparison group. The relative risk ratio cannot make claims about the absolute risk for a particular group (Bottiani et al., 2018; National Research Council, 2002). For example, when examining reading rates, a relative risk ratio of 2.0 suggests that Black students are twice as likely to be identified as having a below basic understanding of reading compared to White students. However, it does not say anything about the rates of the individual groups. In this example, it could mean that 2% of black students are identified as having a below basic understanding of reading and 1% of White students have a below basic understanding of reading. Conversely, it could suggest that 60% of black students are identified as having below basic understanding of reading and 30% of White students have a below basic understanding of reading. A similar problem occurs when using odds ratios because of the use of a reference group (National Research Council, 2002).

Disproportionality defined with reference groups may become unreliable and often difficult to detect when the population is highly homogeneous (e.g., a school whose demographics include 90% of students who identify as Hispanic; Bottiani et al., 2018).

Absolute risk frequencies also referred to as risk indexes for individual groups, has been recommended as an alternative to methods that use reference groups (Bottiani et al., 2018; Losen et al., 2015; National Research Council, 2002). Rather than comparing the risk level of one group to another group, this method subtracts the percentage of individuals identified from the total percentage of students from that group (Losen et al., 2015). Therefore, when using absolute risk frequencies, inferences can be drawn about disproportionality in individual groups better than when using risk ratios (Bottiani et al., 2018). However, relative risk ratios or odds ratios are most commonly used to describe disproportionality in identifying SEB risk (Boneshefski & Runge, 2014). Therefore, most of the information described in the following compares outcomes of minority groups to White students.

Disproportionality in ODRs

The majority of research on disproportional use of ODRs is with Black and White students (Bradshaw et al., 2010b; Girvan et al., 2017; Kaufman et al., 2010; Skiba, Michael, Nardo, & Peterson, 2002). Black students receive significantly more ODRs compared to white students, even when controlling for variables that may impact the disproportional use of ODRs (e.g., free and reduced lunch, grade point average, and school-, classroom-, and individual-level behavior data; Girvan et al., 2017; Martella et al., 2010; Roque, 2010; Skiba et al., 2002). In general, Black students are more likely to receive ODRs for subjective reasons (e.g., disrespect and defiance) whereas their White students are more to receive ODRs for objective reasons (e.g., fighting and vandalism; Girvan et al., 2017; Skiba et al., 2002, 2011; Smolkowski, Girvan, McIntosh, Nese, & Horner, 2016). For example, Girvan and colleagues (2017) found that race account for an additional 1.5 to 3 times the variability in subjective ODRs compared to objective ODRs. Smolkowski and colleagues (2016) also found that Black students were more likely to receive subjective ODRs (OR ranged from 1.25 to 1.73).

A racial match between the teacher and student did not reduce the risk of Black students receiving an ODR (Bradshaw et al., 2010b). Black male students had the highest likelihood of receiving an ODR if their teacher was also Black (adjusted OR = 0.58). Data regarding the disproportionality of ODRs with other ethnic or racial minority groups (e.g., Hispanic, Asian, and Native American/American Indian students) are mixed or less apparent. Hispanic students receive more ODRs than non-Hispanic White students, whereas other studies show that Hispanic students receive fewer ODRs than non-Hispanic White students (Rocque, 2010; Skiba et al., 2011; Whitford & Levine-Donnerstein, 2014).

There are few studies being conducted on the rate of ODRs with Native American and Asian students because of relatively small sample sizes and a focus on other ethnic minority groups (Bradshaw et la., 2010; Skiba et al., 2011; Whitford & Levine-Donnerstein, 2014). Of the studies conducted with Native Americans, several studies have found that Native Americans are overrepresented in the number of ODRs received compared to White students (Wallace, Goodkind, Wallance, & Bachman, 2008; Whitford & Levine-Donnerstein, 2014), while other studies show that Native-American students receive similar or more ODRs than Black students (Brown, 2014; Skiba, Peterson, & Williams, 1997). Asian students have been found to be less likely to receive ODRs compared to other ethnic minority groups (Wallace et al., 2008; Whitford & Levine-Donnerstein, 2014).

Disproportionality in Nominations

The lack research surrounding teacher referral and identification is A significant limitation in understanding disproportionality with teacher nominations for SEB risk. Research on the identification of SEB problems often comes from referrals for special education (e.g., National Research Council, 2002). Although data on referrals for special education (e.g., emotional disturbance and other health impairment) are not equivalent to systematic universal screening methods using teacher nominations, several issues should be mentioned. First, disproportionate number of ethnic minorities have been referred for special education since the inception of special education in the 1970s (National Research Council, 2002). In the category of emotional disturbance, Black students are consistently identified at higher rates compared to White students (Gravois & Rosenfield, 2006; National Research Council, 2002; Zhang, Katsiyannis, Ju, & Robers, 2014). The National Research Council (2002) found that black students were 59% (OR = 1.59) more likely to be identified as having an emotional disturbance compared to white students.

Research has found mixed results with Native American students. Sometimes they are identified with higher rates or equal rates of emotional disturbance compared to White students (National Research Council, 2002; Zhang, Katsiyannis, Ju, & Robers, 2014). Hispanic and Asian students were less likely to be identified with an emotional disturbance compared to White students (National Research Council, 2002; Zhang, Katsiyannis, Ju, & Robers, 2014). These results may be due to the highly subjective nature of teacher identification of SEB risk. Research has found that teachers identify SEB risk in their students through intuition and functional impairment rather specific symptomology (Green et al., 2017). The subjectivity within teacher identification has led researchers to suggest that BBRS may be a superior method for identifying SEB risk because each student is rated on the same criterion (Dowdy et al., 2010).

Disproportionality in Rating Scales

Empirical studies and theoretical manuscripts have examined disproportionality and the use of BBRS as universal screening tools for SEB functioning. Some researchers suggest that the use of BBRS as universal screening tools eliminate or reduce the disproportionality of minority students identified with SEB risk (Raines et al., 2012). Dever, Raines, Dowdy, and Hostutler (2016) examined students that would be identified as at-risk for SEB problems through a self-reported BBRS compared with individuals that were already receiving services for special education and found that Black students were, (a) overrepresented in the groups that did not receive special education, (b) identified as at-risk for SEB problems on the BBRS and received special education services, and (c) not identified as at-risk for SEB problems on the BBRS. This research showed a mismatch between individuals self-reported levels of SEB problems and individuals receiving services for SEB problems.

Given the high correlation between behavioral and academic functioning (Hinshaw, 1992), it would be expected that students would show similar levels of SEB risk across racial/ethnic groups in special education. The use of BBRS as a universal screening tool should reduce the mismatch between symptomology and treatment (Dever et al., 2016). White students rated themselves as having more SEB problems than Black students (Dever et al., 2013), but Splett and colleagues (2018) found that race/ethnicity did not predict being identified by teachers as at-risk for SEB problems by the BBRS tool or by the school's identification method. It may be that there are differences in rates of actual SEB functioning and teacher perceptions or school identification of SEB functioning, but the same biases that are present in the referral processes may also be present in BBRS used in universal screening (Splett et al., 2018).

Differences have been found when comparing racial/ethnic differences on rating scales that are not used for universal screening purposes. For example, Lau and colleagues (2004) found that parents of White children reported higher externalizing and internalizing problems compared to minority youths. Teachers reported fewer internalizing problems for Black adolescents and fewer externalizing problems for Asian individuals, but there were no differences between groups on self-reports of behaviors. Other research has found that when controlling for classroom behaviors, Asian students are rated as having higher levels of externalizing behaviors compared to Black and White students (Mason, Gunersel, & Ney, 2014).

Explanations for Disproportionality

Several possible explanations have been posed to understand the disproportionality in identifying students at-risk for SEB problems including, (a) a mismatch between values of students and teachers/schools, (b) lower socioeconomic status (SES) of minority students, (c) higher rates of behavior problems for Black students, and (d) implicit and/or explicit bias (Bradshaw et al., 2010b; Dever et al., 2013; Girvan et al., 2017; Skiba et al., 2011; Smolkowski et al., 2016; Townsend, 2000; Wallace et al., 2008; Wu et al., 1982). I will discuss each of these explanations below.

Mismatch

Data from a national dataset found that 72-80% of non-white students were instructed by White teachers (McGrady & Reynolds, 2013). Thus, many students will experience a cultural mismatch and this mismatch leads to disproportionate number of ethnic minorities (excluding Asians) identified with SEB problems (Townsend, 2000). Cultural minority students may have different expectations on what is and is not acceptable behavior as determined by the teacher or school at-large, which can lead to perceptions by teachers that students are acting in an unacceptable manner without the student realizing how their behaviors are perceived (e.g., working on an assignment while standing rather than sitting in a seat; Townsend, 2000)). The frequent cultural mismatch has led toward recent efforts to improve culturally responsive teaching to improve student behavior (Larson, Pas, Bradshaw, Rosenburg, & Day-Vines, 2018).

Bradshaw et al. (2016b) attempted to elucidate the discrepancy in cultural mismatch and disciplinary practices, but still found that Black male students were more likely to receive an ODR when their teacher was also Black. The researchers acknowledged that race is not equivalent to culture, and that the racial/ethnic disproportionalities in discipline cannot be explained by a mismatch between race alone. When considering intersectionality and the wide variety of experiences and identities that shape a student’s culture, analysis of any single factor in isolation would provide an incomplete picture of student functioning (Crenshaw, 1993; McCall, 2005). However, this research does indicate that although calls to increase the number of minority educators is a worthy cause, it would not explain the disproportionate number of minority groups identified with SEB problems (McIntosh, Girvan, Horner, & Smolkowski, 2014). In addition, this does not explain why Black male students in particular, compared to other racial/ethnic minority groups, are not identified with the same rates of SEB problems.

SES

Another explanation the disproportionality in SEB risk identification is that minorities are more likely to come from household with lower SES, and research has consistently shown the negative effects of growing up in these environments (e.g., exposure to lead, alcohol, or tobacco and less cognitively stimulating environments; National Research Council, 2002). In fact, by 24 months of age, children from higher SES families know about 450 words whereas children from lower SES families know about 300 words (Fernald, Marchman, & Weisleder, 2013). Prevalence rates for behavior problems were nearly 30% for preschool-age children from low SES families compared to 3% to 6% of preschool-age children from higher SES families (Qi & Kaiser, 2003). However, at the middle and high school level SES has not been shown to be related to SEB functioning (Dever et al., 2013). In separate studies, researchers found disproportionate identification of SEB problems in racial/ethnic minority students even when accounting for SES (Skiba et al., 2002; Wallace et al., 2008).

Higher Rates of Behavior Problems

A third explanation for the disproportionate number of minorities identified with SEB problems is that these groups display higher rates of problems than their White counterparts. This may be the most logical answer to the question of disproportionality in that certain cultural groups are identified with more SEB problems because those groups have higher rates of SEB problems. However, research has shown that minority students, specifically Black male students, display the same levels of SEB problems as their White counterparts in the classroom yet receive harsher punishments (Bradshaw et al., 2010b; McCarthy & Hoge, 1987). As such, differences in the identification of SEB functioning has been theorized to be a result of cultural biases, whether implicit or explicit (Smolkowski et al., 2016; Townsend, 2000).

Bias

Lastly, research has documented both explicit and implicit bias and their roles in how individuals perceive and interact with other individuals (Girvan, Deason, & Borgida, 2015; Smolkowski et al., 2016). Explicit biases are the overt expressions of prejudice or discrimination; these are consciously held beliefs that one group is superior to others (McIntosh et al., 2014). In schools, teachers may have lower academic expectations for Black students compared to White students (Ferguson, 2003). For example, researchers found that teachers had more positive expectations for European American students than Latinx ($d = 0.23$; 95% CI [0.10; 0.37]) and Black students ($d = 0.24$; 95% CI [0.19; 0.27]; Tenenbaum & Ruck, 2007). A similar study of a nationally representative sample found that Non-White teachers' views of White students do not vary from White teachers' views of White students, but Non-White teachers' views of Non-White students' academics and behavior were slightly more positive compared to White teachers' views of Non-White students (McGrady & Reynolds, 2013). These studies were not able to distinguish between conscious and unconscious biases.

Implicit biases are automatic unconscious thoughts that result in behaviors that are discriminatory (McIntosh et al., 2014). Implicit biases occur when an individual lacks the information needed to make an unbiased decision (e.g., interpretation of talking in class as noncompliance when the student is asking a classmate for help) or when the individual lacks the resources to make an unbiased decision (e.g., the decision has to be made quickly or the individual is fatigued) (Pearson, Dovidio, & Gaertner, 2009;

Smolkowski et al., 2016). Research has shown that Black students receive significantly more ODRs for subjective reasons (e.g., insubordination and noncompliance), while White students tend to receive ODRs for objective reasons (e.g., truancy and fighting; Girvan et al., 2017; Skiba et al., 2011; Smolkowski et al., 2016; Wallace et al., 2008). In contrast, Asians are less likely to be identified with SEB problem even though research has found they may have more internalizing behavior problems compared to White students (Lorenzo, Frost, Reinherz, 2000). The differential rates in SEB risk identification may result from teacher expectations of subgroups of students or the manner in which subgroups of individuals display behaviors are different than teacher expectations (Gupta, Szymanski, & Leung, 2011; Lau et al., 2007; Townsend, 2000). Interventions to reduce implicit bias are difficult because the decisions occur unconsciously (McIntosh et al., 2014). Therefore, it would be prudent to identify the behaviors on BBRS that relate to the disproportionate identification of individuals with SEB problems.

Synthesis

Disproportionality can be defined as either the overrepresentation or underrepresentation across groups for a given construct. Disproportionality is usually portrayed in the literature through risk ratios or odds ratios, which compare groups against a reference group, typically White students. Across ODRs, teacher nominations, and rating scales, Black students are more likely to be identified with SEB risk. Research regarding other ethnic and racial groups is inconclusive, with the exception of Asian students. Asian students are less likely to be identified with SEB risk compared to White students. Researchers have provided four main explanations for disproportionality in SEB risk identification, (a) a mismatch between values of students and teachers/schools, (b) increased rates of problem behaviors in minority populations, (c) lower socioeconomic status (SES) of minority students, and (d) implicit and/or explicit bias. Research has not shown the cultural mismatch, minority students have increased rates in problem behaviors, or lower SES as fully explaining disproportionality in SEB risk identification. Research has shown that implicit and explicit biases impact people's perceptions and interactions with other individuals. To identify biases in how perceptions of student behavior may lead to disproportionality in SEB risk identification, an examination of how individuals respond to questions related to SEB risk is warranted.

Theoretical Framework for and Identifying Bias in Screening

Rating scales are frequently used to collect SEB data because each student is evaluated on the same criteria (Dowdy et al., 2010). A theoretical gold standard for using rating scales would be a universally agreed upon operational definition of each survey item with similar definitions on frequencies and intensities that are related to the provided response options for each rater, and deviation from this would result in biased responding (Snow, Cook Lin, Morgan, & Magaziner, 2005). However, each respondent has a unique way of understanding the question and terminology, recalling information, and formulating a response to the rating scale (Jobe, 2003). Information processing models have been developed to understand the steps a respondent uses to understand the question being asked and the cognitive processes performed to develop an answer to the question (Jobe, 2003; Snow et al., 2005; Tourangeau & Rasinski, 1988).

Information Processing Theory and Bias

Tourangeau and Rasinski (1988) developed a four-step information processing model to understand how individuals complete rating scales based off their individual attitudes. In the first step, the individual comprehends and interprets the question (Tourangeau & Rasinski, 1988). In SEB assessment, the individual uses long-term memory to retrieve existing schema of the behavior (Fazio, Sanbonmatsu, Powell, & Kardes, 1986). For example, the individual may represent worry as both verbal and nonverbal expressions of anxieties. The individual may interpret worry with a specific facial feature or physiological response in other individuals. Next, the individual uses their episodic memory of the comprehended and interpreted behavior to identify events. For example, a teacher may recall specific times when a student was worried, or the teacher may recall a previously established summary of a student's worrisome behavior.

Some behaviors may be easier to remember, especially when they occur frequently or with greater intensity (Jobe, 1996; Schwarz, 1999). For example, temper outburst might be easier to recall than impulsiveness because of the intensity of the behavior. In the third step, the individual uses a judgment process from the recalled events or previously established summary to scale their attitude. In this step, the individual places meaning or weight to the events recalled from memory, typically using some integration method like adding or averaging, to formulate a decision about the frequency of a behavior (Tourangeau & Rasinski, 1988). Individuals use heuristics in the integration process to estimate a behavioral frequency. For example, if an individual is determining the frequency of a behavior that has occurred in the past month, they may remember the number of events that took place in the past week and use this number to estimate the frequency of the event in the past month.

In the final stage, the individual reports their judgment about the frequency of a behavior onto the provided response options. In addition, before the individual provides a response, the respondent may edit their judgment by ensuring that the frequency is consistent with their previous responses (Tourangeau & Rasinski, 1988). For example, the individual has an internal judgment of the frequency of a worry and determines that it occurs *Sometimes.* The individual may also check to make sure that their response of *Sometimes* matches the frequency of the behaviors previously rated.

Information Processing and Subjectivity

In the information processing model described here, subjectivity within a response can occur at all four steps, which can produce differential responding. In the first step of comprehending or interpreting the question, there is an inherent subjective nature in what the rater perceives the behavior to look like (Peshkin, 1988; Weber, 2003). Each rater has their own representation of the behavior and this may or may not align with the test developer or other raters. In addition, the behavior may be displayed differently across different groups of individuals (Townsend, 2000). Therefore, the less an individual understands the behavior and how these behaviors are displayed across different groups of individuals, the greater the likelihood of biased responding (Townsend, 2000).

The events or information retrieved from memory are susceptible to subjective bias as well. The frequency of behaviors that occur irregularly can be overestimated (Jobe, 1996) and behaviors that are more intense are more easily remembered (Schwartz, 1999). Therefore, behaviors that are harder to remember are more susceptible to biased responding. In the third step, the individual scales their memories of the behavior. For SEB risk identification, the individual estimates the behavior onto a frequency scale. Making a judgment on the frequency of a behavior can be difficult and people tend to use heuristics during decision-making (Stone et al., 2000). For example, an individual may recall that a behavior occurred two times in the past week, then the individual estimates that the behavior occurred eight times in the past month. In this case, the estimation of the frequency of a behavior is susceptible to the recency of the behavior.

Finally, in the judgment process, the individual uses a method of integrating the behavioral events. Respondents are asked to provide their perceptions of the frequency of a behavior, but intensity and duration are used to determine frequency (Stone et al., 2000). Therefore, the more the individual uses estimation techniques or integration methods that include factors other than frequency (e.g., intensity and duration) then the more the ratings are susceptible to biased responding. In the final step, the individual places their judgment regarding the frequency of the behavior onto the provided response options. Biased responded at this step would occur if the respondent's perception of the frequency of a behavior does not match the provided response options. The individual may go through an editing process before assigning a rating. Therefore, the greater the mismatch between the internal judgment of a behavior and the provided response options, the greater possibility there is for biased responding. Overall, an information processing model may help explain why differential responding occurs across subgroups of individuals.

Identifying Bias in Assessment

Evaluations of disproportionality in rating scale measurement can be done by analyzing the overall proportion of individuals identified on a construct of interest. Alternatively, statistical methods have been used to evaluate the response patterns between groups on the individual items present on rating scales. In test development, measures may be given to an expert panel that identifies items that might be biased and should be removed from the measure. Once the measure is completed and used with different groups of individuals, it may be assumed that there is measurement invariance. That is, the measure and items are functioning equally for different groups of individuals (e.g., the item is measuring the construct for males the same way in which the item is measuring the construct for females; Kim & Yoon, 2011). However, measurement invariance cannot and should not be assumed, especially in the identification of SEB functioning in which minorities are identified as at-risk at disproportionate rates. Comparisons between groups on SEB functioning may not be valid if analyses have not been conducted on measurement invariance (Borsboom, 2006). Therefore, the purpose of checking for measurement invariance is to ensure fairness and equality in testing across groups of individuals (Zumbo, 2007).

Measurement Invariance

Measurement invariance is defined as the variability in the scores obtained are a function of the latent construct of interest and is unrelated to other characteristics, such as group membership (Kim & Yoon, 2011; Zumbo, 2007). For example, the observed scores on individual items in a SEB measure are a result of that individual's SEB functioning and is not influenced by factors not associated with the latent construct (e.g., race/ethnicity, biological sex, and SES). Measurement invariance can be tested through confirmatory factory analysis (CFA) and item response theory (IRT) (Kim & Yoon, 2011).

Kim and Yoon (2011) compared CFA and differential item functioning (DIF) in IRT using ordered categorical data (e.g., Likert scale items). In their study, the researchers found that DIF performed better than CFA in identifying measurement invariance in both true positive and false positive rates. Therefore, DIF within an IRT framework will be used for the purposes of this study in identifying measurement invariance in screening assessment measures. A brief overview of IRT is provided before discussing DIF.

Item Response Theory

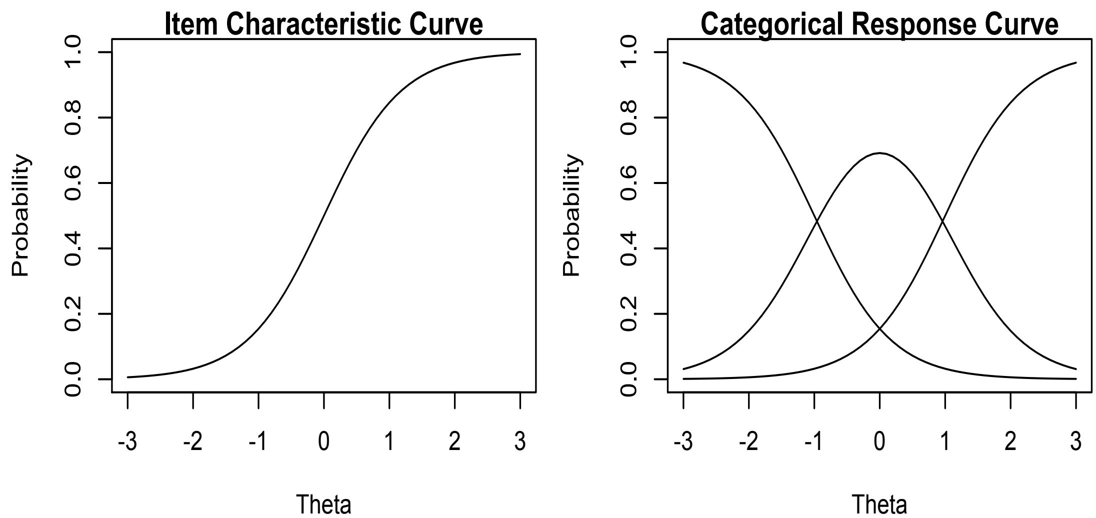
Classical test theory (CTT) has been the predominant method for constructing, analyzing, and scoring psychological measurements (Bock, 1997; Reise, Ainsworth, Haviland, 2005). Item response theory (IRT) is an alternative approach to measurement that offers a variety of advantages over CTT. IRT models the relationship between the latent trait of the individual with the performance of an item used to measure the latent trait (Nguyen, Han, Kim, & Chan, 2014). Therefore, each item has a different probability of providing a particular response for each individual (Reise et al., 2005). In addition, the item and individual are estimated on the same scale, theta (θ), which follows a z-score distribution with a mean of 0 and standard deviation of 1 (Cappelleri, Lundy, & Hays, 2014). By measuring the latent trait of the individual and the items of a measure on the same scale, reliability can be calculated (i.e., through standard error) at different levels of the latent trait. This is different from CTT in which Cronbach's alpha is calculated for the measure and is the same for each item and person. Lastly, the items on the measure are independent of the individuals that complete the measure (Nguyen et al., 2014). In CTT, an individual's observed score is dependent upon the items administered. For example, a student in third grade that completed items with single-digit addition problems would have a higher raw score than if that same student completed items with algebra problems, but their underlying math knowledge is the same. In IRT, the individual's estimate of math knowledge would be similar no matter which items were administered. These advantages describe the three fundamental characteristics of IRT: (a) item response functions (IRF), (2) information functions, and (c) invariance (Reise et al., 2005).

Item Response Function

The item response function (IRF) describes the relationship between the underlying latent trait of an individual and the probability of providing a particular response (Reise et al., 2005). For dichotomous data, the IRF is presented as an item characteristic curve (ICC; Figure 1). The ICC displays the continuum of the latent trait on the x-axis (represented by theta or θ) and the probability of providing a correct response on the y-axis (Reeve & Fayer, 2005). Item difficulty or severity level (*b*) and item discrimination or slope parameter (*a*) are needed to estimate the probability of the individual responding in a particular manner. Item difficulty for dichotomous is defined as the inflection point of the ICC. This occurs at a theta value of 0 in Figure 1. The item discrimination parameter describes how well an item can differentiate between individuals with different levels of the underlying latent trait (Reeve & Fayer, 2005). The difficulty parameter is defined by the slope of the ICC at the inflection point. The steeper the curve or the higher the discrimination parameter the greater the item can differentiate between individuals of varying levels of the latent trait (Nguyen et al., 2014).

Figure 1

The Item Characteristic Curve (ICC) for Dichotomous Data and the Categorical Response Curve for a Polytomous Item with Four Response Options



Polytomous data (i.e., greater than two response options) functions similarly to dichotomous data. The ICC in dichotomous data is referred to as categorical response curves (CRC; Figure 1) and are plotted for each response category (Nyugen et al., 2014; Reeve & Fayer, 2005). For polytomous data, the difficulty parameter is the point in which the individual is more likely to respond to one category over another. There are *k-1* difficulty parameters, in which *k* is the number of response categories. The discrimination parameter can be understood by the amount of overlap between CRC, with less overlap indicating greater discrimination. Item difficulty and discrimination can be used to provide item and scale information or item reliability at different levels of the latent trait.

Item and Scale Information

Item and scale information describe the reliability across the latent trait (Reise et al., 2005). For example, an easy math problem is better at differentiating between individuals with low levels of math knowledge latent trait. A test with lots of easy math items will be useful in differentiating between people with low levels of math skill. In IRT, reliability is described in terms of information and standard errors, in which the information function is the inverse of the standard error function. In the previous example, easy math items would provide more information at low levels the math latent trait and less information at high levels of math latent trait. Conversely, there would be smaller standard errors at low levels of the latent trait and larger standard errors at high levels of the latent trait. Therefore, items provide information based off the difficulty and discrimination of the item. Items with greater discrimination provide more information at the difficulty parameter of the item.

Each item information parameter can be added together to describe the scale information. The scale information function describes the reliability of a measure across the latent trait (Reise et al., 2005). Therefore, the reliability of a scale differs depending on the latent trait of the individual, whereas in CTT, reliability is the same for each individual (Reise et al., 2005).

Invariance

Invariance describes two concepts in IRT, (a) an individual's latent trait can be estimated from items with known IRFs, even if those items come from different measure, and (b) the IRF does not depend on the particular group in which the individual belongs (e.g., latent trait level, race/ethnicity, biological sex, or socioeconomic status) (Reise et al., 2005). This means that item parameters (i.e., discrimination and difficulty) remain stable, even when those items are administered to different groups of individuals (Reeve & Fayer, 2005). However, measures should be evaluated to determine if this assumption is met. Items are described to have differential item functioning (DIF) when there are differences in difficulty or discrimination between groups.

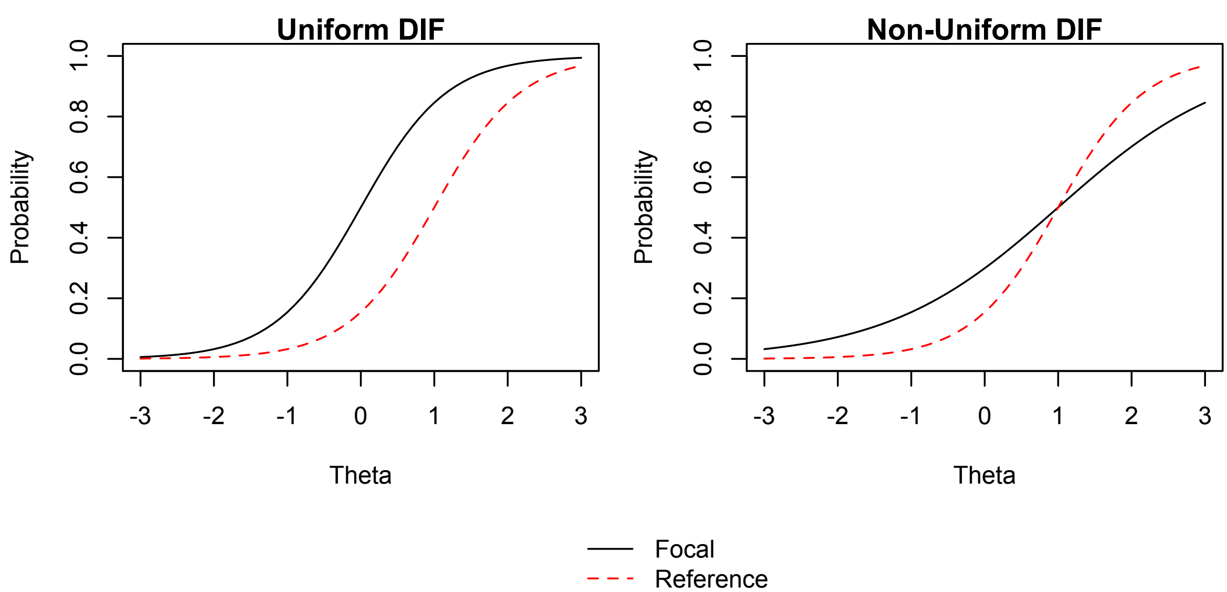
Differential Item Functioning (DIF)

DIF can be used to identify measurement invariance between a focal group and reference group (e.g., White vs minority or male vs female; Zumbo, 2007). DIF can be the result of item impact or item bias (Zumbo, 2007). Item impact refers to actual differences between groups that result in different response patterns on a measure. For example, individuals with ADHD will respond differently than individuals without ADHD on items related to attention, and these differences can be attributed to actual differences between the groups. On the other hand, item bias occurs when measurement invariance is not met due to some characteristic of the person responding or the item (Zumbo, 2007). For example, teachers might rate Black students differently on items of aggression compared to White students due to underlying perceptions and interpretations of the items rather than actual difference between Black and White students.

In the IRT framework, DIF is established when the ICC of the focal group is significantly different than the ICC of the reference group (Figure 2; de Ayala, 2009; Zumbo, 2007). Therefore, DIF can occur when one item is more difficult for one group compared to the other (i.e., uniform DIF) or when the discrimination parameter is different from one group to the other (i.e., non-uniform DIF; Zumbo, 2007). When an item differs only on difficulty, (i.e., uniform DIF) it can be said that an item is more difficult for one group compared to the other group, even when individuals in the different groups have the same level of the latent trait. In practical terms for SEB functioning, uniform DIF would represent consistently lower ratings on items with DIF, even when the individuals have the same overall SEB functioning. When an item differs only on discrimination it can be said that there is an interaction between ability and group membership (Zumbo, 2007). For example, an item with non-uniform DIF would be more difficult at low levels of the latent trait, but easier at high levels of the latent trait for the focal group compared to the reference group (Kristjansson, Aylesworth, McDowell, & Zumbo, 2005).

Figure 2

Item Characteristic Curves Showing Uniform and Non-uniform differential item functioning (DIF) for Dichotomous Data

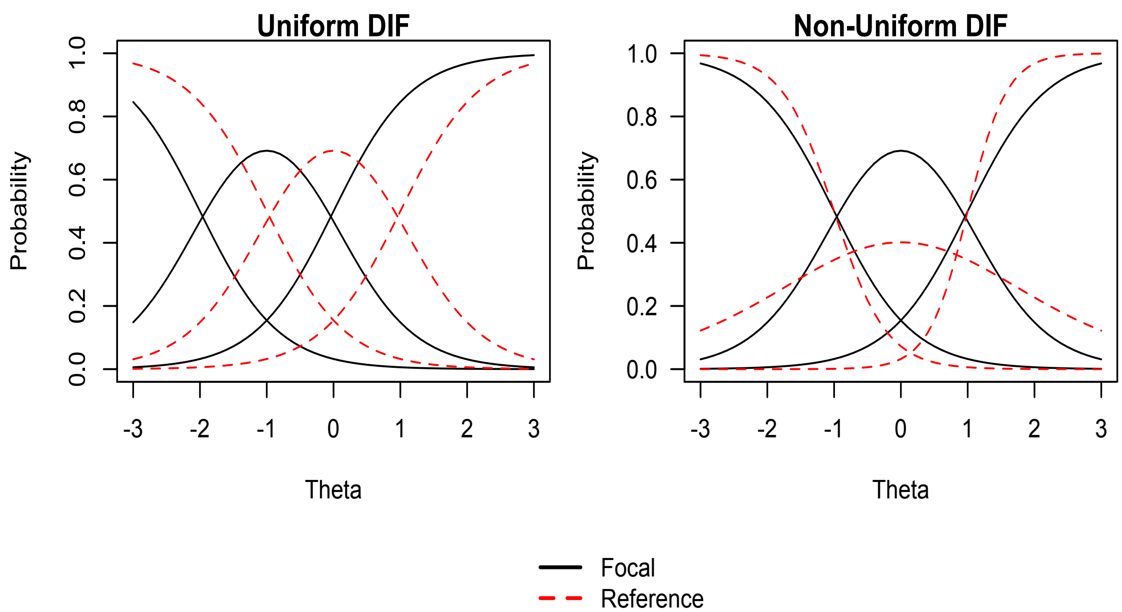


Identifying and interpreting DIF becomes more difficult with ordered categorical data (e.g., Likert scale items) because DIF has to be conducted for each response option (Kristjansson et al., 2005). These types of data are commonly collected when evaluating

SEB functioning (e.g., BESS; Kamphaus & Reynolds 2015), and therefore, these types of analyses should be explored. Similar to DIF with dichotomous data, there can be uniform or non-uniform DIF with ordered categorical data (Figure 3; Kristjansson et al., 2005). The meaning of uniform and non-uniform DIF is the same when using ordered categorical data and dichotomous data; however, the interpretation can look different. An extension of the definition used for uniform DIF with dichotomous data to ordered categorical data would indicate that the one group has lower probabilities of responding to k versus $k – 1$, where k represents the response, when the two groups have the same latent trait level. For example, on an item displaying uniform DIF with response options of *Never, Sometimes, Often,* and *Almost Always* for the questions "I am sad," a group is more likely to respond to *Often* instead of *Sometimes*, even though the two groups have the same overall level of SEB functioning. In non-uniform DIF it can be said that one group's probability of responding k versus $k – 1,$ changes with latent trait level of the individual. Taking the same example, individuals with low levels of a latent trait in the focal group may have higher probabilities of responding to *Never* rather than *Sometimes*, but have higher probabilities of responding to *Often* rather than *Sometimes* at high level of the latent trait.

Figure 3

Categorical Response Curves Displaying Uniform and Non-uniform differential item functioning (DIF) for Polytomous Data



Note. Uniform DIF displays a shift in the difficulty parameter only. Non-uniform DIF displays a difference in discrimination parameter only. All the difficulty parameters remain the same as noted by the location of the intersection points between response categories, but the curves change.

Synthesis

Information processing theory can be used to understand the process that teachers go through when providing responses on perceptions of behaviors displayed by their students. A four-step process proposed by Tourangeau & Rasinski (1988) includes: (1) comprehension of the question, (2) recall of events related to the behavior, (3) estimating the frequency based off the events recalled, and (4) placing their estimated frequency onto the provided response options. However, this four-step process is susceptible to biased responded at each step and these four steps can be used to understand bias in assessment. Bias in assessment can be due to item impact or item bias. Item impact refers to actual difference between groups on a latent construct. Item bias refers differential responding due to some underlying characteristic of the person responding or the item. Both of these biased responding result in measurement invariance or differences in response patters across different groups of individuals. DIF within IRT can be used to assess measurement invariance. With order categorical data (e.g., Likert style questions), CRC are compared between the focal group (e.g., Black males) and the reference group (e.g., White males). DIF can be described as uniform, an item is more difficult for one group compared to another group even when those groups have similar levels of the latent trait, or non-uniform, there is an interaction between item difficulty and theta level of the individual. Research has begun investigating DIF with universal screening measures for SEB risk.

Differential Item Functioning with Brief Behavioral Rating Scales

Three studies have conducted descriptive DIF for universal SEB risk screeners. Dowdy and colleagues (2011) compared DIF for individuals with limited English proficiency with students considered English proficient on the teacher version of the BESS (Kamphaus & Reynolds, 2007), and found that most items did not display DIF. However, these results should be taken with caution as the number of participants in the study were low for detecting DIF using IRT. That is, 142 students were in the focal groups and 110 were in the reference group, with recommendations that a minimum of 500 individuals are in each group (Embretson & Reise, 2000; Reeves & Fayers, 2005; Reise & Yu, 1990; Revicki et al., 2014).

Lambert and colleagues (2018) evaluated the Emotional and Behavioral Screener (EBS; Cullinan & Epstein, 2013) for DIF across Black, White, and Hispanic students. The researchers found that overall impact of DIF was small to negligible. Only two items for males and three items for females displayed DIF (all R^2 values were less than .035). However, the study only examined teacher ratings for first grade students. The EBS is aligned with the emotional disturbance special education category, and it may be that increased DIF would be present for students in different age categories.

Finally, Schatschneider and colleagues (2014) evaluated DIF by gender, age group, and special education status on the Student Risk Screening Scale (SRSS; Drummond, 1994). The researchers found DIF on each item for each comparison, but the effect sizes for these results were generally negligible when comparing boys and girls (Cohen's $d < .12$ for all items). When larger effect sizes were found, the researchers noted that the results of DIF were due to item impact rather than item bias (Schatschneider et al., 2014). For example, when comparing students by special education status, students in special education had much lower ratings on academic achievement (Cohen's $d = -.88$). The result of finding significant DIF for each item and comparison made may be due to the large number of participants in the study (Kim, Cohen, Alagoz, & Kim, 2007; Meade, 2010). Therefore, when determining the significance of DIF with large samples, effect sizes may be more appropriate than p values for interpretation of results (Meade, 2010).

Several effect sizes have been described in the literature that relate to the effect of DIF on ratings of an individual item and the test as a whole (Meade, 2010). These methods include visual methods (Kim et al., 2007) and statistical methods (Meade, 2010;

Zumbo, 1999). Zumbo (1999) describes a method using ordinal logistic regression to obtain R^2 values as estimates of effect size that combines uniform and non-uniform DIF detection. However, other researchers have noted that not enough research is available to determine the efficacy of this measure as an indicator of DIF effect size (Kim et al., 2007). Kim and colleagues (2007), suggest using a visual method with descriptive information. Visual information can be displayed in graphs including the difference in response functions between the focal and reference group, the impact on item score, the impact of DIF on the total score, and the difference in difficulty and discrimination for the focal and reference group. They suggested that visual inspection can aid in interpretability, and with descriptive information tied to the visuals would provide additional context.

Meade (2010) described a method of calculating effect sizes at the item and test level that standardize the information that is provided through visual analysis. At the item level, the average expected score difference between the focal and reference group can be compared. This is similar to a graph displaying the difference in difficulty and discrimination and the impact on item score between the focal and reference group. Similarly, the expected test scores differences between the focal group and reference group can be compared. This is similar to a graph display if the sum of DIF across all items between the focal group and reference group. Therefore, when estimating the effects of DIF, graphs with descriptive information described by Kim and colleagues (2007) and calculated effect sizes described by Meade (2010) are both beneficial in understanding the effects of DIF at the item and test level.

Synthesis

DIF have used to identify difference in response patterns between groups on individual items on BBRS (Kim & Yoon, 2011; Kristjansson et al., 2005; Zumbo, 2007). Previous research in DIF with BBRS tools that can be used for universal screening have focused on identifying problematic items rather than explaining the reasoning for the DIF (Dowdy et al., 2011; Lambert et al., 2018; Schatschneider et al., 2014). In addition, these studies have not examined DIF that may occur across different racial/ethnic groups or the interaction of race/ethnicity and biological sex across a range of age groups.

Study Purpose

Further research is needed to understand how BBRS function across individuals by race/ethnicity and the interaction of biological sex and race/ethnicity. Research indicates that when using disciplinary approaches (e.g., ODRs) or informal referrals to identify SEB risk, minority groups are identified at disproportional rates. BBRS used as a universal screening method have been suggested as a means to reduce disproportionality in SEB risk identification because each student is being evaluated on the same criteria (Dowdy et al., 2013; Raines et al., 2012). However, in an information processing model of rating scale responding, teachers rate SEB functioning on BBRS that is susceptible to subjective responding. It is unclear if students from different groups are being assessed similarly by BBRS used for SEB universal screening purposes. Therefore, the purpose of this study is to evaluate SEB risk identification by answering four questions:

1. What are the frequencies of SEB risk by race/ethnicity and the interaction of biological sex and race/ethnicity?
2. What items contribute toward disproportionality in SEB risk identification?

3. To what extent are there any trends in items that display DIF by group membership?
4. How well can DIF on a SEB BBRS be predicted by examining how individuals complete rating scales or by the subjectivity within each item?

CHAPTER III: METHOD

This chapter focuses on the research method of the current project. Study one examined questions one, two, and three, which are related to disproportionality. Study two attempted to answer question four, explaining the measurement invariance in study one.

Study 1: Questions 1, 2, and 3

Study one examined if a disproportionate number of minority students were identified as at-risk for SEB problems in universal screening practices. Items on a brief behavioral rating scale (BBRS) tool used for universal screening of SEB problems were examined to determine the rate of disproportionate identification of SEB risk. Analyses examined measurement invariance on a teacher completed SEB rating scale. Study one attempted to answer the following research questions.

1. What are the frequencies of SEB risk by race/ethnicity and the interaction of biological sex and race/ethnicity?
2. What items contribute toward disproportionality in SEB risk identification?
3. To what extent are there any trends in items that display DIF by group membership?

Participants

The current study used data collected for research purposes by *FastBridge Learning*, a company that provides formative assessments in reading, math, and behavior to help schools facilitate their MTSS process. This dataset included teacher behavior screening data from 11,525 students from 176 schools across the United States. Measurement invariance was conducted on data gathered from the Social, Academic, and

Emotional Behavior Risk Screener (SAEBRS) Teacher Rating Scale (TRS).

In this sample, 51.3% of the students were male and 48.7% of the students were female (see Table 1 for participant demographics). Fifty-three percent of the population was White, 26.7% Black, 11.0% Hispanic, 5.6% multiracial, 2.5% Asian, and 0.8% were Native American. The ages ranged from 4 to 18 (M =10.4, SD = 2.3), and the grades ranged from kindergarten to 12th (M = 4.9, SD = 2.4). Five percent of the student received special education services.

Table 1

Descriptive Statistics for the Social, Academic, and Emotional Behavior Risk Screener – Teacher Rating Scale (N = 11,524)

Variable	Data
Age	M = 10.4 SD = 2.3
Grade	M = 4.9 SD = 2.4
Biological Sex	
Male	51.3%
Female	48.7%
Race/Ethnicity	
Asian	2.5%
Black	26.7%
Hispanic	11.0%
Multiracial	5.6%
Native American	0.8%
White	53.4%
Receiving Special Education Services	5.1%

Measure

For the purposes of this study, only the *Emotional Behavior* (EB) scale of the SAEBRS-TRS was used. The EB scale was used because it has shown higher variability in responding compared to the other scales (Kilgus, Eklund, von der Embse, Taylor, Sims, 2016). The EB subscale measures internal states of functioning including regulating emotions, adapting to changes, and responding to stressful events. The EB scale is related to internalizing behavior problems and resilience. Lower scores on the SAEBRS-TRS EB scale are associated with increased risk. Scores between 0 and 16 are considered at-risk and scores between 17 and 21 are considered not at-risk. Previous studies have found acceptable levels of reliability, validity, sensitivity, and specificity (e.g., Kilgus, Chafouleas, & Riley-Tillman, 2013; Kilgus et al., 2016; Kilgus, Sims, von der Embse, & Riley-Tillman, 2015). Internal consistency for the EB scale of the SAEBRS-TRS for these data was .84, which is above the criterion of Cronbach alpha > .80 for low-stakes decision-making (Salvia et al., 2016).

Procedures

Deidentified data were collected by the test publisher, *FastBridge Learning,* and provided with permission to the lead author. The SAEBRS-TRS was used as a universal screening tool to assess SEB risk. Data from each school were only provided if the school screened at least 80% of their student population, which matches universal screening procedures. Data were collected from January 2016 to November 2017, at one or more time periods (i.e., fall, winter, and spring over two school years). Data were only used for the first time the individual was screened. Therefore, each individual only had results from one teacher screening measure.

Extant student demographic data were provided by the school or school district, including data related to race and ethnicity, age, gender, grade, and special education status. As such, some individuals had missing demographic information.

Analyses

The first two research questions were analyzed with OR and DIF. Each is described below.

Research Question #1: Frequencies

First, analyses were conducted to examine if the SAEBRS-TRS identified a disproportionate number of students at-risk based on their relative and absolute risk frequencies. Research has suggested that both methods of calculating disproportionality should be reported (McIntosh, Ellwood, McCall, & Girvan, 2018). Risk status was calculated to align with EB raw scores (i.e., at-risk on the SAEBRS-TRS EB scale = raw score of 0-16 and not at-risk = raw score of 17-21). Relative risk ratios were calculated by dividing the proportion of individuals identified by the proportion identified in the reference group. This was calculated with the following formula:

$$Risk\ Ratio = \frac{\left(\frac{\#\ subgroup\ identified\ at\ risk}{\#\ subgroup\ \ in\ sample}\right)}{\left(\frac{\#\ all\ other\ students\ identified\ at\ risk}{\#\ all\ other\ students\ in\ sample}\right)} \qquad (2)$$

For the purpose of this study, all other students of the same biological sex not in the subgroup were used as the reference group. This method allows for risk ratios to be calculated for all groups of individuals (IDEA Data Center, 2014). For example, White female students were compared to all other non-White female students. This method maintains independence of the groups, and is preferable to comparisons that include the focal group in the reference group (IDEA Data Center, 2014). Absolute frequencies were determined with the following equation:

$$Absolute\ Risk = \frac{Subgroup\ \#\ identified\ at{-}risk}{Subgroup\ \#\ in\ total\ sample} \qquad (3)$$

The absolute frequencies were calculated for race/ethnicity and the interaction of biological sex and race/ethnicity. Absolute risk allows for comparisons to be made between subgroups in study, and also allow for comparisons to be made across studies (Losen et al., 2015).

Research Question #2: Differential Item Functioning (DIF)

A unidimensional graded response model was fit to the entire sample before conducting DIF. This was done to determine discrimination and threshold parameters for the entire sample. The discrimination parameter describes how well an item differentiates between individuals with different levels of the underlying latent trait (Reeve & Fayer, 2005). Each threshold parameter represents the point at which the individual is more likely endorse one category over another. There are *k-1* threshold parameters, in which *k* is the number of response categories. Therefore, there were three threshold parameters for each item in the current study. Lastly, a test information curve was created to display the region of the latent trait continuum within which maximum measurement precision can be obtained.

Differential item functioning (DIF) within IRT was used to evaluate if racial/ethnic groups and the interaction of race/ethnicity and biological sex demonstrated measurement invariance. DIF was conducted for each item on the SAEBRS-TRS EB subscale by comparing response patterns by race/ethnicity and the interaction of race/ethnicity and biological sex. Researchers have suggested that 500 responses are needed to responses to obtain stable parameters (Embreston & Reise, 2000; Reeves & Fayers, 2005; Revicki et al., 2014; Tsutakawa & Johnson, 1990). The standard errors may be too large to be considered stable with less than 500 responses. An inspection of the standard errors was done for groups with less than 500 responses to determine if they could be included in the current study.

The groups that had less than 500 individuals were Asian students (n = 291), male students with multiple races (n = 324), female students with multiple races (n = 318), and

Native American students (n = 87). The standard errors for the discrimination parameters were, on average, eight times larger for these groups compared to the total sample's standard errors (see Appendix A). This indicates that with additional students, the standard errors would become much smaller, and the current items cannot be considered stable for these groups in this study. Therefore, DIF analyses could not be conducted for Asian students as a whole, Native American students as a whole, and for both male and female students of multiple races and ethnicities.

DIF compares how one group responds to that of a reference group. For the purposes of this study, all other students were used as the reference group, similar to calculation of risk ratios. When conducting DIF on the interaction of race/ethnicity and biological sex, only students from the same biological sex as the focal group were used as the reference group. For example, all White students were compared to all non-White students and Hispanic females were compared with all other non-Hispanic female students.

A unidimensional method of DIF was used because only one scale of the SAEBRS-TRS was used (i.e., EB subscale). Kristjansson and colleagues (2005) describe and compared four methods for identifying DIF with ordered categorical data: (1) the Mantel, (2) the Mantel-Haenszel, (3) logistic discriminant function analysis, and (4) ordinal logistic regression. These methods produce a p-value associated with the probability of the item displaying DIF as the criterion for identifying items (Kristjansson et al., 2005). However, a p-value criterion does not provide information on the significance or meaning of DIF (Borsboom, 2006). In addition, items are likely to display DIF because of the large sample sizes required for running DIF (e.g., minimum of 500 individuals; Embretson & Reise, 2000; Reeves & Fayers, 2005; Reise & Yu, 1990; Revicki et al., 2014). Therefore, the current study used visual methods and effect size estimates when describing DIF.

Effect Sizes

Effect sizes were calculated as described in Meade (2010) and displayed with descriptive information as described by Kim and Yoon (2007) and Choi, Gibbons, and Crane (2011). Two effect sizes were calculated, one at the item level and one at the test level, that are comparable to Cohen's (1988) *d*. The expected score standardized difference (ESSD; Meade, 2010) is an effect size at the item level and was calculated by computing the expected scores for the focal group using the parameters for both the focal group and reference group. Then, the differences in scores were divided by the pooled standard deviation. This is an item level effect size and compares the observed scores with the expected scores. The following equations were used to calculate the ESSD

$$ESSD_i = \frac{\overline{ES}_{\gamma F} - \overline{ES}_{\gamma R}}{SD_{ItemPooled}} \tag{4}$$

where $\overline{ES}_{\gamma F}$ is the mean score for the focal group using the item parameters for the focal group and $\overline{ES}_{\gamma R}$ is the mean expected score for the focal group using the item parameters for the reference group. The item pooled standard deviation was calculated with the following formula

$$SD_{ItemPooled} = \sqrt{\frac{(N_F - 1)SD_{ES(i|\gamma F)} + (N_F - 1)SD_{ES(i|\gamma R)}}{2 \times N_F - 2}} \tag{5}$$

where N_F is the sample size of the focal group. Therefore, the ESSD is provided in standard deviation units. The expected test score standardized difference (ETSSD;

Meade, 2010) is similar to the ESSD, except this effect size at the test level rather than item level. This was calculated with the following formula

$$ETSSD_i = \frac{\overline{ETS}_{\gamma F} - \overline{ETS}_{\gamma R}}{SD_{TestPooled}} \tag{6}$$

where $\overline{ETS}_{\gamma F}$ is the mean test score for the focal group using the item parameters for the focal group and $\overline{ETS}_{\gamma R}$ is the mean expected score for the focal group using the item parameters for the reference group. These effect size estimates and score differences were calculated within the mirt package (Chalmers, 2012) in R (R Core Team, 2018). Cohen's (1988) *d* recommendations for small, medium, and large effect size values (.20, .50, and .80, respectively) were used to interpret these values.

In addition, unstandardized differences in scores were calculated to indicate the expected difference in raw scores at the item level (i.e., signed item difference in the sample [SIDS]) and at the test level (i.e., signed test differences in the sample [STDS]). The SIDS can be understood as the average expected score difference for the focal group when using the item parameters for the focal and reference group (Meade, 2010). The SIDS allows for cancellation when items display non-uniform DIF. For example, if the DIF for the focal group has higher expected scores at lower latent trait levels, but lower expected scores at higher latent trait levels, then DIF can cancel each other out. The STDS is the sum of SIDS and can be understood as the average expected test score differences between the focal and reference group for individuals with the same latent trait level. Similar to the SIDS, the STDS allows for cancellation across items. For example, positive SIDS values cancelled out with negative SIDS scores. The SIDS and STDS values are interpretable through the original terms of the SAEBRS-TRS EB subscale. For example, a SIDS of 0.5 on item 1 for Hispanic students would mean that

Hispanic students would be expected to score 0.5 points higher compared to all non-Hispanic students on item 1, when controlling for the latent variable. Likewise, a STDS of -2 for White students would mean that White students would be expected to score 2 points less than all other non-White students on the EB subscale, when controlling for the latent variable.

Next, five graphs were generated to visualize and complement the numeric effect sizes. Two graphs represent the test level and three at the item level. An expected total score graph displayed the expected score on the EB scale of the SAEBRS across the range of theta estimates for the focal group using the item parameters of the focal group and the focal group responses with the item parameters of the reference group. The test score differences were displayed in a separate graph, which displayed the differences between the test characteristic curves in the expected total score graph previously described. Positive values on this graph represent higher expected scores for the focal group compared to the reference group and negative values represent lower expected scores for the focal group compared to the reference group. This graph also displays where along the range of theta values DIF has the most impact, and is a visual representation of the STDS.

The same graphs were also created at the item level. The expected score graph displays the expected score for the focal and reference group for each item across theta values. The expected score graph display two item response functions (IRF). Both IRFs use the focal groups item responses, but one graph uses the item parameters from the focal group and the other uses the item parameters from the reference group. An item score difference graph was created which displays the difference between the expected item score curves. Positive values represent higher expected scores on the item for the focal group compared to the reference group and negative scores represent lower expected scores on the item for the focal group compared to the reference group. This graph displays where along the range of theta values DIF has the most impact for the item and is a visual representation of the SIDS. Lastly, the categorical response curve (CRC) shows the difference in probability of responding to each category between the focal group and reference group. This graph is a visual representation of the difference between the item parameters for the focal and reference groups (i.e., discrimination and threshold parameters). Overall, when describing the impact of DIF, raw score and effect size differences were used along with their visual representations.

The same procedures were also done when the data were disaggregated by gender. This was done to determine if the interaction of race/ethnicity and gender had an effect on DIF. For example, DIF was analyzed between White male students and all other Non-White male students. Then, the effect sizes for White male students were compared with White female students. Differences in effect sizes would indicate an interaction between race/ethnicity and gender.

Study 2: Question 4

The purpose of Study 2 was to determine if DIF on the EB subscale of the SAEBRS can be explained by underlying characteristics of how teachers perceive the test items (Zumbo, 2007). Specifically, the second study was conducted to determine if overall DIF could be predicted from an information processing theory (IPT) model of rating scale responses or by the perceptions of subjectivity of each item of the SAEBRS-TRS EB subscale. The second study addressed the following research question:

4. How well can DIF on a SEB BBRS be predicted by examining how individuals complete rating scales or by the subjectivity within each item?

The second study sampled teachers to examine the process they consider when completing the SAEBRS-TRS EB subscale items using IPT and their perceptions of the subjectivity of each item on the SAEBRS-TRS EB subscale. A separate set of teachers were sampled in study two from the group of teachers used in study one.

The assumption of invariance in IRT, which states that item parameters are not dependent upon a particular population, allows for a separate sample of teachers to be used in study two (Reise et al., 2005). However, the assumption of invariance does not signify that the item parameters will be the same regardless of sample characteristics. The results of DIF identified in study one were calibrated from a large representative sample of students. The item specific statistical properties (i.e., discrimination and threshold parameters) are expected to hold with other similar samples. Therefore, a sample of teachers were recruited for study two that taught students with similar racial/ethnic and grade distribution as those used in study one. In addition, the teachers recruited in study two were already completing the SAEBRS-TRS as part of their school-based practice, separate from the current study procedures.

Participants

Participants included 48 teachers from Midwest schools that were already using the SAEBRS-TRS (see Table 2 for participant demographics). Eighty percent of the teachers were female and 20% were male. The teachers were 85% White, 4% Black, 6% Hispanic, 2% Asian, and 2% Native American. The teachers taught grades kindergarten through 12 with a mean grade level of 5.5 and standard deviation of 4.2. Ninety percent of teachers taught general education students and 10% taught special education students. The average teacher respondent had been teaching for 12.2 years with a standard deviation of 8.3 years.

Table 2

Descriptive Statistics for the Teachers (n = 48) *In Study 2 of Their Perceptions of the Emotional Behavior Subscale Items of the Social, Academic, Emotional Behavior Risk Scale.*

Variables	Percentage
Biological Sex	
Male	79.2%
Female	20.8%
Race/Ethnicity	
White	85.4%
Black	4.2%
Hispanic	6.3%
Asian	2.1%
Native American	2.1%
Students Taught	
General Education	89.6%
Special Education	10.4%
Grade Taught	M = 5.5 SD = 4.2
Years Teaching	M = 12.2 SD = 8.3

Measure

Based on an exhaustive search of the literature, to our knowledge no current measure has been created that evaluates how individuals respond to questions about their attitudes or perceptions about behaviors that children display. Therefore, a measure was created to evaluate how teachers complete the SAEBRS-TRS EB scale in order to better understand and describe DIF. A search of the literature resulted in multiple articles on information processing theory, with Tourangeau and Rasinski's (1988) four-step process consistent across different models of information processing theory (Jobe, 2003). A measure using the four-step process of information processing theory proposed by Tourangeau and Rasinski (1988) was used to develop a measure to understand the decision making process teachers consider when completing BBRS. When creating a measure, research recommends that a content validation study be completed on a new measure before it is used (McKenzie, Wood, Kotecki, Vlark, & Brey, 1999; Rubio, Berg-Weger, Tebb, Lee, & Rauch, 2003).

Content Validation

Content validity refers to the extent to which a measure captures the full breadth of the construct (e.g., information processing theory; McKenzie et al., 1999). Tourangeau and Rasinski (1988) describe the four-step process as (1) comprehension of the question/behavior, (2) memory of the behavior, (3) decision making of the frequency of the behavior, and (4) formulation of a response onto the provided response options. Questions were developed for each of the four steps. The questions were provided to doctoral level experts in the field of SEB assessment and feedback was provided on wording as well as development of additional questions. A content validation form was created once the final list of questions for each step were finalized (See Appendix C). Five advanced level graduate students in a school psychology were recruited to complete the content validation study, with a minimum of three experts needed (Lynn, 1986).

Each graduate student was provided the instructions to complete the content validation form, which included the purpose of the measure, the rating scale, and example ratings. A brief synopsis of the four-step information processing theory defined by Tourangeau and Rasinski (1988) was provided. Then, each graduate student was asked to indicate which question best represents the particular step, with 1 being the best representation of the step (see Appendix C for the final questions in the Item Response Questionnaire). Respondents marked each question until there were no more questions for that step. Rankings were collected, so that the best question could be identified for each step.

Next, each graduate student marked how confident they were with their ranking, according to the following response options of *Not Confident, Somewhat Confident, Mostly Confident,* and *Very Confident*. Lastly, each graduate student marked how relevant the question was to the step, with response options of *Not Relevant, Somewhat Relevant, Mostly Relevant*, and *Very Relevant*. Respondents followed these steps for each of the four steps in information processing theory. At the very end, the graduate students were asked if any questions were missing from any step.

Data were analyzed by calculating interrater agreement once all five graduate students completed the content validation form. First, rankings were evaluated by inspecting the mean ranking across raters for each item. The question with the lowest average score was determined to be the best representation of the step. Next, the questions with the lowest ranking from each step was inspected for confidence and relevance scores. First, confidence ratings were dichotomized (e.g., *Not Confident* = 0, *Somewhat Confident* = 0, *Mostly Confident* = 1, and *Very Confident* = 1), as described in research (Rubio et al., 2003). Next, the ratings were summed and divided by five, the number of raters. Rankings were considered confident if they had a reliability score greater than .80 as recommended in research (Davis, 1992). The same process was completed for the relevance rating. This process was used to determine the four best questions, with each question representing one step in information processing theory proposed by Tourangeau and Rasinski (1988).

The final four questions are displayed in Table 3, including overall ratings from content validators. The IRQ was used to measure teachers' perceptions of how objectively they rated behaviors on the SAEBRS EB subscale. That is, if teachers were better able to recognize a behavior, used specific or discrete events of the behavior, used less estimation, and did not compare their ratings to previous behaviors, then they would be rating the particular behavior more objectively. As such, the third and fourth questions were reverse scored so that each item was measured on the same scale.

Procedures

Following approval from the Institutional Review Board, the researcher met with graduate students in school psychology to complete the content validation for the information processing measure, as described above. The lead author discussed the purpose of the IRQ and the intended use of the measure. After analyzing content validation ratings, pre-service teachers were solicited to provide feedback in a focus-group style session on their thoughts of the measure before recruiting teachers to complete the IRQ.

Table 3

Information Processing Theory questions for Teachers

Question	Ranking	Confidence	Relevance
I can recognize this behavior when it is displayed.	1.0	1.0	1.0
I use specific/discrete events of this behavior when rating this question.	1.2	1.0	1.0
I rate the frequency of this behavior based off an estimation of how often the student engages in the behavior.	1.2	1.0	1.0
I compare my previous ratings on other behaviors when rating this question	1.4	1.0	1.0

Note. Rankings range from 1 to 5 with 1 being the best item for that step, confidence and relevance range from 0 to 1 with 0 being not confident/relevant and 1 being confident/relevant.

Teachers were recruited from schools that were heterogenous by race and ethnicity. That is, because this measure is focused on identifying DIF as a function of race and ethnicity, the teachers completing the IRQ came from a school with a diverse student population. Once the teachers completed a consent form indicating their willingness to participate in the study, individuals completed the IRQ via an online survey using the Qualtrics survey platform. Each participant was asked to read one item from the EB subscale followed by four items on the IRQ. Each participant then rated each item on the SAEBRS-TRS EB subscale using the IRQ. Following the IRQ, the next page displayed the seven behaviors on the EB subscale of the SAEBRS-TRS. Teachers ranked the questions from one to seven in terms of their perceptions of the subjectivity of the question, with one being the most subjective. Rankings on this section of the measure were a forced choice. The IRQ and rankings of subjectivity of the behaviors were used to explain DIF effect sizes on each item.

Analyses

Before using the IRQ to predict DIF effect sizes, an analysis of the IRQ was conducted. A correlation table was created that compared each item with the total score of the IRQ in order to assess that the IRQ was a unidimensional measure. Once the unidimensionality of the IRQ was established, the scores were used to predict effect sizes of DIF (i.e., ESSD) that were computed in study one.

The ESSD can be interpreted in the same manner as Cohen's *d* (Meade, 2010). Therefore, positive and negative effect sizes can be produced. When combining effect sizes, research has suggested using weights proportional to the inverse of the variance in each study (Hedges 1982). The purpose of using proportional weights is to give more weight to studies with larger sample sizes because larger sample sizes should produce more precise effect sizes (Hedges, 1985). This method is typically used in meta analyses (Hedges & Vevea, 1998). The current study used a single sample of individuals, with each effect size produced from the same total sample. For example, White students were compared to all non-White students.

In addition to the effect sizes being developed from the total sample, IRT has different assumptions that those used in CTT (Reise et al., 2005). The current study established IRF for the EB scale of the SAEBRS that was disaggregated by race and ethnicity and by the interaction of race and ethnicity and biological sex. Once stable item parameters were established (e.g., the IRF is known by calibrating the items), the assumption of invariance is met. For example, the item parameters for Hispanic males and all other non-Hispanic male students are stable. Therefore, the effect size of DIF can be assumed to be stable once IRF are known for each group. Therefore, no weights were used when combining the effect sizes.

The effect sizes were combined using the absolute median effect size. The median effect size was used rather than the mean for each effect size because the median is not influenced by outliers and there were only four effect sizes per item (i.e., each racial/ethnic group produces one effect size per item). In addition, the absolute value of each effect size was used because the current study is interested in explaining the impact of DIF, not the direction of DIF.

The IRQ questions were used to predict the ESSD for each item separately using a micro-macro multilevel model (Bennick, Croon, & Vermunt, 2013; Croon & van Veldhoven, 2007). The micro-macro model is a multilevel model, in which the dependent variable *Y* was measured at the group level (i.e., group level effect sizes of each item based off ratings from teachers on the SAEBRS) and can be predicted by lower level variables (i.e., individual teacher perceptions of behaviors using IRQ) (Croon & van Veldhoven, 2007). This model is in contrast to typical hierarchical linear models in which the dependent variable *Y* is measured at the lower or individual level and is influenced by higher level or group level predictors. The micro-macro model was selected over a linear regression because the linear regression would result in biased regression parameters

(Croon & van Veldhoven, 2007). In addition, the micro-macro model has greater power for detecting individual-level predictor variables compared to a linear regression (Foster-Johnson & Kromrey, 2018). The micro-macro model proposed by Croon and van Veldhoven (2007) uses a two-step approach that identifies the adjusted group means using multi-level modeling followed by an ordinary least square analysis (Foster-Johnson & Kromrey, 2018). The adjusted group mean can be calculated with the following formula:

$$\tilde{\boldsymbol{x}}_g = (1 - w_g)\mu_\xi + w_g\bar{\boldsymbol{x}}_g \tag{6}$$

Where $\tilde{\boldsymbol{x}}_g$ is a vector of adjusted group means, $\bar{\boldsymbol{x}}_g$ is a vector of the observed group means, μ_ξ is the overall mean, and w_g is calculated with the following:

$$w_g = \frac{\sigma_\xi^2}{\sigma_\xi^2 + {}^{\sigma_\xi^2}\!/_{n_g}} \tag{7}$$

The adjusted group mean of the predictor variables were then regressed on the group level *Y* variable. The regression equation can be written as follows:

$$y_g = \beta_0 + \tilde{\boldsymbol{x}}'_g\beta + \epsilon_g \tag{8}$$

Where y_g is the group level outcome variable (i.e., median absolute effect size for each item), $\tilde{\boldsymbol{x}}'_g$ is a vector of the adjust group mean for each question on the IRQ, and ϵ_g is the error term. This method was calculated in R using the lme4 package (Bates, Maechler, Bolker, & Walker, 2015). R^2 values represent the amount of variance in effect size of DIF that can be explained by the IRQ.

Next, the subjective rankings of the SAEBRS EB scale from each teacher was used to explain the rank order of DIF effect sizes. DIF as measured by the ESSD was ranked from least to most DIF. A Kendall's tau was used to correlate absolute median effect size with teachers' rank ordering of subjectivity.

CHAPTER IV: RESULTS

Chapter 4 discusses the results of the research. The results are organized according to Study 1 (Research Questions 1 and 2) and Study 2 (Research Questions 3 and 4).

Study 1

Study 1 was conducted to determine overall disproportionality in risk identification on the EB scale of the SAEBRS-TRS. In addition, the study used DIF within IRT to identify the items that contributed to the disproportionality in risk identification.

Question 1 – Absolute Risk and Risk Ratios

Overall, 32.1% of the total sample was identified as at-risk on the EB scale of the SAEBRS. Absolute risk of individuals identified by race and ethnicity and the interaction of race and ethnicity and biological sex varied by group (Table 4). Black students and Native-American students had risk ratios greater than 1, indicating higher identification for emotional risk on the SAEBRS compared to their proportion in the sample. Asian students, Hispanic students, students with multiple races/ethnicities, and White students had risk ratios less than 1, meaning they were less likely to be identified with emotional risk on the SAEBRS compared to their proportion in the sample. The same risk trend was observed when disaggregated by biological sex; however, absolute risk was higher for males and lower for females.

Table 4

Absolute Risk and Risk Ratios by Race and Ethnicity and the Interaction of Biological Sex and Race and Ethnicity Race/Ethnicity

	All		Female		Male	
	AR	RR	AR	RR	AR	RR
Asian	19.93%	0.63	19.44%	0.62	20.42%	0.65
Black	46.81%	1.82	39.95%	1.33	53.14%	1.90
Hispanic	26.16%	0.82	22.92%	0.72	29.28%	0.93
Multiple	30.19%	0.96	26.42%	0.84	33.33%	1.06
Native American	48.28%	1.54	44.19%	1.41	52.27%	1.67
White	25.18%	0.65	22.06%	0.64	28.00%	0.86
Total	31.4%		27.2%		35.2%	

Note. AR = absolute risk. RR = risk ratio, all other individuals not in the focal group are used as the reference group for risk ratios.

Question 2 – Differential Item Functioning

A graded response model was fit to the entire sample of students before conducting the DIF analyses. The discrimination and difficulty thresholds are displayed in Table 5. The discrimination parameters for all items were large, and mostly within the range of good discrimination (i.e., 0.8 to 2.5; de Ayala, 2008). Only the item measuring sadness had a discrimination parameter outside the range at 2.97. Second, all but one difficulty threshold was in the negative range. These results suggest the measure may provide more information for individuals with lower emotional behavior functioning compared to individuals with average or above average functioning. This can be seen in the information plot displayed in Figure 4.

Table 5

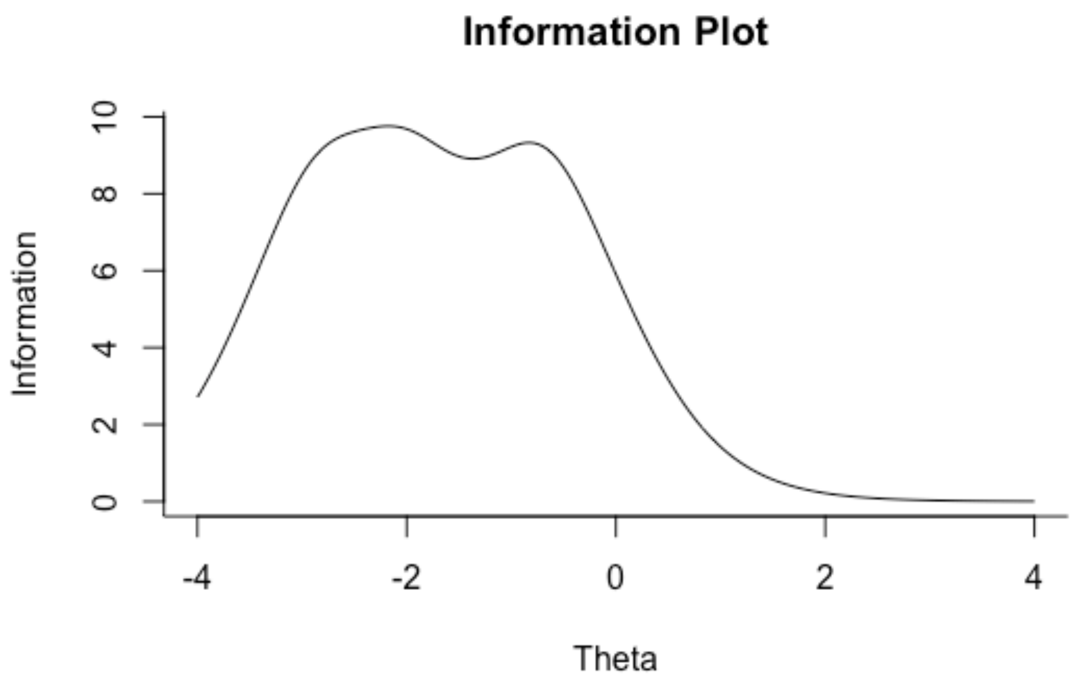
Discrimination and Threshold Parameters for the Total Sample (N = 11,524)

	SAEBRS-TRS Emotional Behavior subscale			
		Difficulty Thresholds		
Item	Discrimination	*0 – 1*	*1 – 2*	*2 – 3*
Adaptability	1.91	-2.43	-1.01	0.01
	(.04)	(.04)	(.02)	(.02)
Difficulty rebounding from setbacks	2.06	-2.24	-1.51	-0.36
	(.04)	(.04)	(.03)	(.02)
Nervousness	2.38	-3.45	-2.40	-1.17
	(.07)	(.09)	(.04)	(.02)
Positive attitude	2.10	-2.93	-1.25	-0.08
	(.05)	(.06)	(.02)	(.02)
Sadness	2.97	-2.91	-2.01	-0.70
	(.08)	(.06)	(.03)	(.02)
Withdrawal	2.47	-2.68	-1.85	-0.63
	(.06)	(.05)	(.03)	(.02)
Worry	1.85	-3.26	-2.12	-0.44
	(.04)	(.07)	(.04)	(.02)

Note. SAEBRS-TRS = Social, Academic, & Emotional Behavior Risk Screener Teachers Rating Scale. Standard errors are displayed in parentheses.

Figure 4

Test Information Plot for the Entire Sample of the Social, Academic, & Emotional Behavior Risk Screener Teachers Rating Scale Emotional Behavior Subscale.

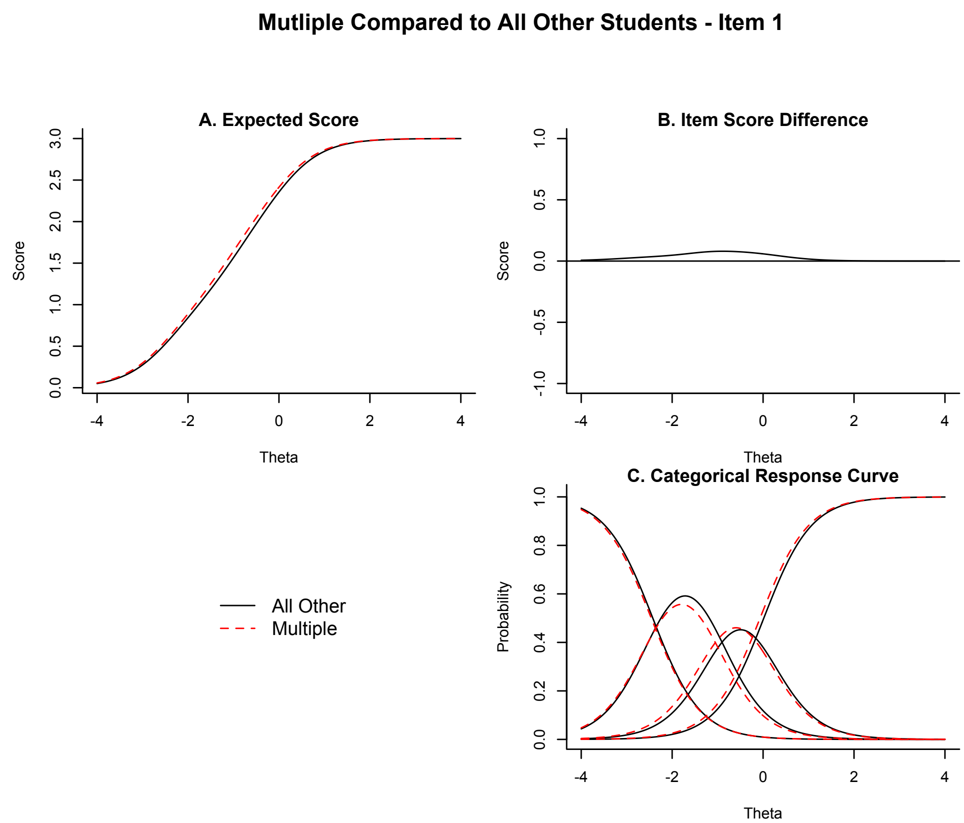


Next, a series of DIF analyses were conducted on the different groups of individuals to identify the impact of DIF across race/ethnicity and the interaction of race/ethnicity and biological sex. For these analyses all other students of the same biological sex were used as the reference group.

DIF. DIF was smallest for students with multiple races and ethnicities and largest for Black students (Table 6). For students with multiple races and ethnicities, the impact of DIF was small to negligible with ESSD ranging from 0.00 to 0.18. A visual analysis of the impact of DIF for students with multiple races and ethnicities for item 1, *Adaptability*, can be seen in Figure 5. The graph shows that the two item response functions are overlapping, indicating less DIF (Figure 5.A). The graphs for the other items are similar to item one's graph and are displayed in Appendix D. Using Cohen's *d* recommendations (1989), all effect sizes were less than the criterion for a small effect (i.e., $d = 0.20$).

Figure 5

Graphs Displaying the Expected Score Difference and Categorical Response Curve between the Focal Group (i.e., Students with Multiple Races and Ethnicities) and the Reference Group (i.e., all other students) on Item One 'Adaptability'



Note. Graph A displays the item characteristic curve across the range of theta values. Graph B displays the difference between the test characteristic curves in graph A. Positive values in graph B represent higher estimated scores for students with multiple races and ethnicities. Graph B also indicates where along the theta range the expected score difference on item one is largest. Graph C displays the categorical response curve for the two groups. This graph plots the probability of responding to category *k* across theta values.

Table 6

Score and Effect Size Indices for Differential Item Functioning Comparing Each Subgroup with All Other Students

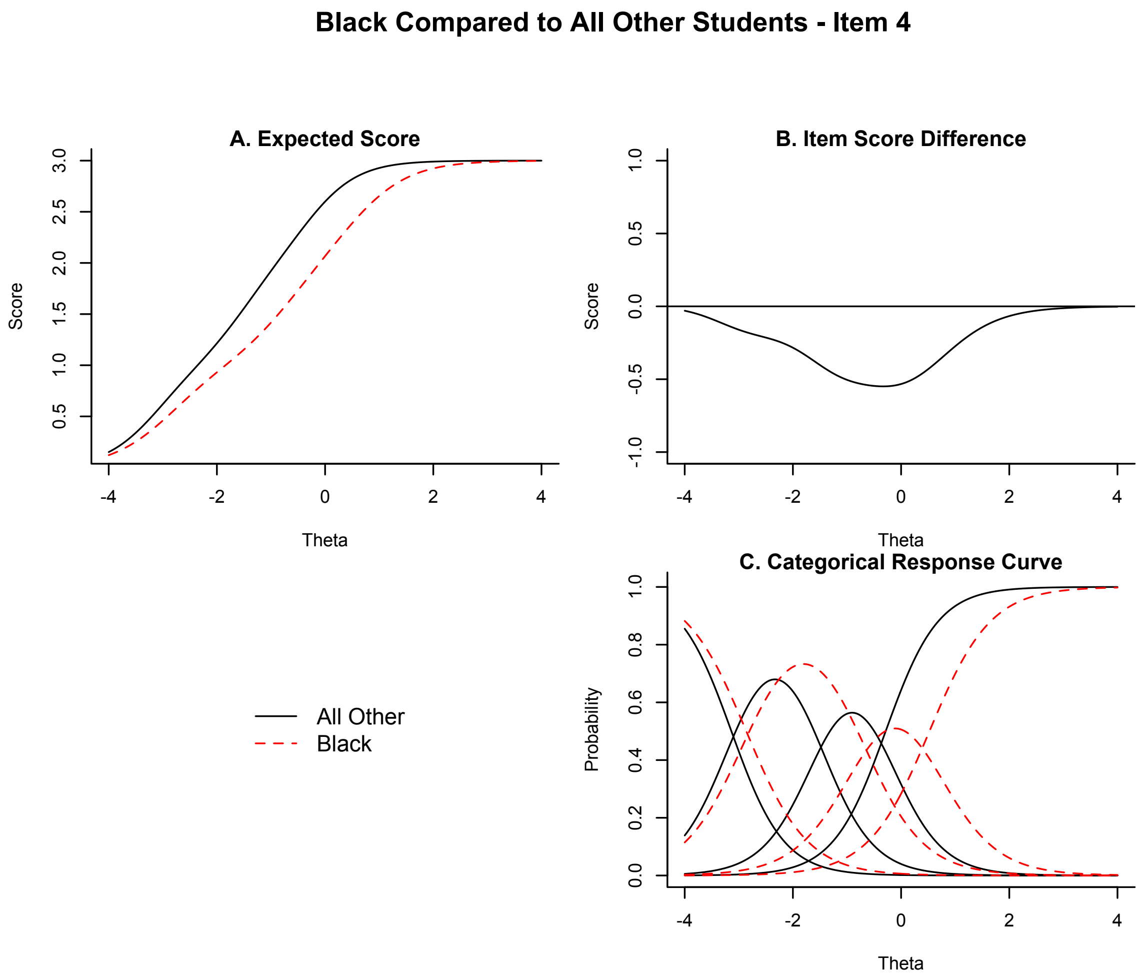
	Black *n* = 3,072		Hispanic *n* = 1,273		Multiple *n* = 646		White *n* = 6155	
Item	SIDS	ESSD	SIDS	ESSD	SIDS	ESSD	SIDS	ESSD
Adaptability	-0.45	-0.82	0.04	0.07	0.05	0.09	0.29	0.53
Difficulty rebounding from setbacks	-0.39	-0.75	0.12	0.25	0.02	0.04	0.21	0.40
Nervousness	-0.05	-0.18	-0.01	-0.05	0.03	0.10	0.04	0.14
Positive Attitude	-0.43	-0.87	0.07	0.14	0.00	0.00	0.28	0.54
Sadness	-0.14	-0.35	0.03	0.09	-0.03	-0.07	0.09	0.21
Withdrawal	-0.19	-0.44	0.02	0.05	0.01	0.02	0.12	0.28
Worry	-0.04	-0.11	0.05	0.13	0.07	0.18	0.00	-0.01
STDS	-1.68		0.33		0.14		1.03	
ETSSD		-0.56		0.11		0.05		0.33

Note. SIDS = signed item difference in the sample. The average difference in expected scores on that item compared to all other students. ESSD = expected score standardized difference. Cohen's *d* for expected score differences. STDS = signed test differences in the sample. The sum of SIDS across items. ETSSD = expected test score standardized difference. Cohen's *d* for expected test score differences. Negative numbers indicate lower scores, which indicates higher risk.

The effect sizes for Black students varied more compared to other student groups, with ESSDs ranging from -0.11 to -0.87, which corresponds to small to large effect sizes. Black students were the only group that had large effect sizes. *Positive attitude*, *Adaptability*, and *Difficulty rebounding from setbacks* had ESSDs of -0.87, -0.82, and -0.75. A visual analysis of item four, *Positive attitude*, displaying a large effect size is displayed in Figure 6. On this item, Black students were expected to score 0.43 raw score points lower (SIDS for Black students on *Positive attitude* in Table 6) compared to all other students. The difference between item response function curves (Figure 6A) can help in interpreting the meaning of these effect sizes. For example, on *Positive attitude,* Black students have to have an emotional behavior latent trait of 0.95 standard deviations above the mean to obtain the same expected score as all other students with an expected emotional behavior latent trait at the mean. Graph B in Figure 6 shows that the impact of DIF is largest for individuals with estimated theta values of emotional behavior between 2 and 1.

Figure 6

Graphs Displaying the Expected Score Difference and Categorical Response Curve Between Black Students and All Other Students on Item 4, Positive Attitude.



Effect sizes for White students ranged from medium to negligible. Effect sizes for Hispanic students were relatively small, with only one item having an effect size above the criterion of small (i.e., *Difficulty rebounding from setbacks d* = 0.25). Black students were the only group that had negative effect sizes for each item. Hispanic, White, and students with multiple races and ethnicities all only had one item with a negative effect size. For each of these groups, the negative effect size occurred on different items, but all of the negative effect sizes were negligible (i.e., *d* was smaller than -0.03 for all the groups).

The effects of DIF resulted in negligible to small effects at the test level for students with multiple races and ethnicities and Hispanic students (*d* = 0.05 and 0.11, respectively). Test level differential functioning was moderate for Black and White students (*d* = -0.56 and 0.33, respectively). Black students were the only group that had a negative test level effect size. A visual representation at the test level effect size for moderate effects is displayed in Figure 7. This figure compares White students to all other students and their expected test score difference. All other test level graphs are displayed in Appendix D.

Figure 7

Graphs Displaying the Difference in Expected Test Scores Between White Students and All Other Students

Note. Graph A displays the test characteristic curve across the range of theta values for the focal group (i.e., White students) and the reference group (i.e., all other students). Graph B displays the difference between the test characteristic curves in graph A. Positive values in graph B represent higher estimated scores for White students compared to all other students, which relates to lower risk. Graph B also indicates where along the theta range the differences in expected test scores is largest. For White students, the test score differences are largest between the theta ranges of -2 to 0.

Interaction of Race/Ethnicity and Gender

A series of DIF analyses were also conducted for each racial and ethnic group disaggregated by biological sex to examine if effect sizes of DIF differed by biological sex. In general, effect sizes were larger for males compared to females (Table 7). For each racial and ethnic group, the difference in effect sizes between males and females on each item ranged from 0.19 on *Nervousness* for Black students to 0.0 on *Positive attitude* for White students. On average the effect size of DIF differed by 0.08 between males and females. Only *Worry* for females changed from being a slightly negative effect size in the total sample to a positive effect size; however, both effect sizes were negligible. In general, the effect size for DIF by race and ethnicity when disaggregated by biological sex did not change the criterion of interpretation of effect size. For example, *Adaptability* had large effect sizes for Black males and females, negligible effect size for Hispanic males and females, and medium effect sizes for White males and females.

Table 7

Effect Score Standardized Difference for Males and Females by Race and Ethnicity

	Black		Hispanic		White	
	Male	Female	Male	Female	Male	Female
Item	ESSD	ESSD	ESSD	ESSD	ESSD	ESSD
Adaptability	-0.92	-0.76	0.11	0.02	0.55	0.53
Difficulty rebounding from setbacks	-0.83	-0.70	0.23	0.27	0.45	0.37
Nervousness	-0.27	-0.08	-0.04	-0.05	0.22	0.05
Positive Attitude	-0.91	-0.85	0.17	0.06	0.55	0.55
Sadness	-0.40	-0.28	0.11	0.06	0.24	0.18
Withdrawal	-0.47	-0.40	0.06	0.04	0.31	0.25
Worry	-0.15	-0.07	0.09	0.17	0.06	-0.08
ETSSD	-0.62	-0.50	0.12	0.09	0.37	0..29

Note. ESSD = expected score standardized difference. ETSSD = expected test score standardized difference.

Question 3 – Trends in DIF

Overall, effect sizes of DIF were larger for males than females at the test level and item level. However, the trends in the effect sizes were similar for the total sample and when disaggregated by biological sex. There were two categories of effect sizes for DIF. The first group included three items, which had absolute median effect sizes above 0.31 (i.e., *Adaptability*, *Difficulty rebounding from setbacks*, and *Positive attitude*; Table 8) and the second category had absolute median effect sizes below 0.20 (i.e., *Nervousness*, *Sadness*, *Withdrawal*, and *Worry*). The first group (i.e., median effect sizes above 0.31), included the only two positively worded items and the second group (i.e., median effect sizes below 0.20), included only negatively worded items.

Table 8

Absolute Median Expected Test Score Standardized Difference Effect Sizes Across Race/Ethnicity for Each Item on the EB Subscale of the Social, Academic, Behavioral Risk Screener

Item	Absolute Median Effect Size
Adaptability	0.31
Difficulty rebounding from setbacks	0.33
Nervousness	0.12
Positive attitude	0.34
Sadness	0.15
Withdrawal	0.17
Worry	0.12

Results Study 2

Study two was conducted with a separate set of teachers to provide additional context as to why DIF could occur on the EB scale of the SAEBRS using teacher perceptions of student behaviors. A micro-macro multilevel model (Croon & van Veldhoven, 2007) was fit to determine if the process that teachers' consider when completing a rating scale, as defined through information processing theory (IPT; Tourangeau & Rasinski, 1988), could be used to predict the DIF effect sizes found in study one. This multilevel model was used because it describes a method for explaining group level outcomes with individual level predictors. In the first step, the adjusted group means were calculated in a multilevel model (Table 9). Adjusted group mean question one of the IRQ indicated that teachers 'Somewhat disagree' with being able to recognize the behavior in the classroom for five out of the seven behaviors (i.e., adjusted group means below 3). The results of the IRQ on question three indicated that teachers use an estimation methods when rating frequency of a behaviors on all items of the SAEBRS EB subscale except for *Adaptability*. The adjusted group means of question four of the IRQ indicated that teachers tended to compare their ratings on other questions before providing a response on all behaviors of the SAEBRS EB subscale.

Next, the adjusted group means from each question of the IRQ were used in an OLS regression to predict the group level median effect sizes for each question. The IRQ did not significantly predict median DIF effect sizes on the EB subscale of the SAEBRS, $R^2 = 0.64$, $F(4, 2) = 3.662$, $p = 0.23$ (Table 10). In addition, none of the individual predictors were significant.

Table 9

Adjusted Group Means for the Social, Emotional, Academic Behavior Risk Screener Emotional Behavior Subscale on Each Question of the Item Response Questionnaire (IRQ)

	IRQ Question			
Item	Q1	Q2	Q3	Q4
Adaptability	2.86	3.00	2.94	2.23
Difficulty rebounding from setbacks	3.05	3.13	2.88	2.22
Nervousness	2.41	2.89	2.78	2.14
Positive attitude	3.69	3.28	3.16	2.29
Sadness	2.99	3.02	2.93	2.05
Withdrawal	2.88	2.97	2.84	2.14
Worry	2.29	2.77	2.86	2.13

Table 10

Results of Ordinary Least Squares Regression Using Adjusted Group Means (n = 7)

Predictor	β	SE	*t*	*p* value
Intercept	-2.34	2.41	-0.97	0.43
Question 1	-0.03	0.41	-0.09	0.94
Question 2	0.39	0.97	0.40	0.73
Question 3	-0.02	0.60	-0.03	0.98
Question 4	0.71	0.54	1.31	0.32
R^2	0.64			0.23

A Kendall's tau correlation was conducted to determine if rankings of teachers' perceptions of EB subscale item subjectivity were correlated with the rank ordering of DIF effect sizes. The hypothesis that rank ordering of perceptions of subjectivity would be positively correlated with rank ordering of effect sizes was not confirmed in this study (τ_b = -.15; $p < .05$). Teachers perceived items that displayed larger DIF effect sizes as less subjective and items that displayed smaller DIF effect sizes as more subjective.

CHAPTER V: DISCUSSION

The current study was conducted to examine the measure properties of the SAEBRS EB subscale, including disproportionality between racial/ethnic groups according to teacher ratings of student behaviors and the interaction of race/ethnicity and biological sex on these ratings. Disproportionality was investigated in three stages: (1) identifying the proportion of risk status by racial/ethnic group based off the raw scores, (2) examining item level measurement invariance between racial/ethnic groups when controlling for the latent variable, and (3) explaining DIF on the EB subscale items by examining the process teachers go through when rating behaviors and through teachers' perceptions of subjectivity.

Risk and Measure Properties

Overall, the SAEBRS EB subscale identified 31.2% of students as at-risk, which is greater than the 20-25% in multitiered systems of support models (Severson et al., 2007). Universal SEB screening is typically used within a school's multitiered system of support, and one piece in identifying students that would benefit from universal, at-risk, or indicated interventions. With over 30% of individuals identified as at-risk in the current study, schools are unlikely to have the resources to provide Tier 2 or Tier 3 supports to all students identified (Kilgus & Eklund, 2016). In addition, risk identification by race/ethnicity largely supported previous research that Black and Native American student had the highest risk ratios and White and Asian students had the lowest risk ratios (Ready & Wright, 2011; Redding, 2019; Tenenbaum & Ruck, 2007). Schools may vary greatly in their racially and ethnic makeup, which will likely result different number of students identified as at-risk. This will exacerbate the load on some schools and the ability to provide services to students that are identified.

A unidimensional graded response model was also fit to the total sample to examine the functioning of the EB subscale. All of the items demonstrated good sensitivity to changes in emotional behavior, with discrimination parameters between 1.85 and 2.97. The discrimination parameter indicated that the questions and response options provide a lot of information at the threshold values, or the point at which an individual is more likely to respond to k versus $k - 1$. The threshold values or difficulty parameters, which were on a *z*-score scale with a mean of zero and standard deviation of one, were all negative, which indicated that the EB subscale of the SAEBRS was better able to differentiate between students with estimated latent traits below theta values of 0. Taken together, the negative threshold parameters and high discrimination parameters indicated that the SAEBRS provided better information for individuals below estimated theta values of 0. The information function indicated the EB subscale provided the most information for individuals below an estimated theta value of 0, which is appropriate for a risk screener because most students will not be at-risk, and therefore, information is not needed for students in the average to above average range. Rather, information is needed to differentiate between not at-risk and at-risk individuals. From an interpretation and use argument (Kane, 2013), the current study supports previous research indicating the SAEBRS is best used for universal screening purposes as it includes items that can distinguish between those individuals with and without risk (Kilgus et al., 2015).

Differential Item Functioning

Measurement invariance was examined between the response patterns of racial and ethnic groups through effect sizes of DIF developed by Meade (2010). The measurement invariance method was preferred over likelihood ratio methods because significant DIF may be the results of large sample sizes that were required to conduct the analysis, and p-values associated with DIF did not inform about the size or significance of DIF. Two effect sizes were used at the item level and two effect sizes were used at the test level.

The current study partially supported previous research documenting differences in teacher perceptions on internalizing behavior problems based off race and ethnicity (Lambert et al., 2018). Lambert and colleagues (2018) found the effect sizes of DIF were negligible to small by race and ethnicity; however, the current study found that effect sizes were medium to large for Black and White students. Specifically, Black students were the only racial or ethnic group that had negative effect sizes for all items, and were the only group that had large effect sizes for some items (i.e., *Adaptability*, *Difficulty rebounding from setbacks,* and *Positive attitude* had ESSD effect sizes of -0.82, -0.75, and -0.87 respectively). These items were also the largest positive effect sizes for White students, with ESSD effect sizes ranging from 0.40 to 0.53. It may be that difference were found between the studies because the current study included students from a wider range of grades. The current study included students from kindergarten through twelfth grades, whereas the Lambert and colleagues (2018) study only used first grade students.

The effect sizes on all of the items were small to negligible for Hispanic students and students with multiple races or ethnicities. In the current study, all other students not in the focal group were in the reference group. Therefore, the reference group included the negative response pattern for Black students and the positive response pattern for White students when doing the DIF analyses for Hispanic students and students with multiple races and ethnicities. This may explain the small effect sizes found with Hispanic students and students with multiple races and ethnicities.

Only the effect sizes for Black and White students were greater than the criterion for a small effect size (i.e., > .20). Negative implicit biases toward Black students and positive implicit biases toward White students have been shown in previous research (Downey & Pribish, 2004; McGrady & Reynolds, 2013). Implicit bias may impact the behavioral expectations teachers have based on the race of the student, which influences how they rate their students. For example, research has found that preservice teachers rated the description of a Black student as more likely to engage in problem behaviors and for the behaviors to remain stable over time (Kunesh & Noltemeyer, 2019). However, other research has found that teachers do not differ in their perceptions of positive and negative traits between Black and White students when teachers were explicitly asked to rate the percentage of students that display specific traits by the race of the student (Chang & Demyen, 2007; Chang & Sue, 2003). It may be that when explicitly asked to rate students based on student race, there are little differences in perceptions by race (i.e., participant bias), but when race is masked (e.g., through the use of a stereotypical racial name) or when bias has to be examined at the group level (i.e., the current study), then implicit bias becomes more apparent. The current study examined measurement invariance through DIF by analyzing the response patterns of thousands of students that were universally screened. At the individual level, a lower rating on the EB subscale items cannot be distinguished from the student's actual emotional behavior latent trait. However, when examining different races and ethnicities across the entire sample, measurement invariance through DIF can be detected.

Item and Scale Functioning by Biological Sex

The current study supports previous research that greater number of males were rated at risk on the EB subscale of the SAEBRS (Dever, Raines, Dowdy, & Hostulter, 2016; Young et al., 2010). In the current study, 35.2% of males were identified as at-risk for emotional behavior problems, compared to 27.2% of females. Research has found greater internalizing behavior risk on universal risk screening measures (Young et al., 2010), even though females have higher rates of diagnoses of internalizing problems starting in adolescence (e.g., depression and anxiety; Bor et al., 2014; Perou et al., 2013). The results of DIF based on biological sex indicated that for the majority of DIF analyses, the effect sizes were slightly larger for males. Specifically, the intersectionality of being Black and male resulted in greater DIF effect sizes. The large percentage of Black males that were identified as at-risk in this study (i.e., 53.14%) was impacted by the three items large DIF effect sizes (i.e., *Adaptability*, *Difficulty rebounding from setbacks*, and Positive *Attitude*). When combined, these three items result in an excepted raw score of 1.43 points less for Black students compared to all other students. On the SAEBRS EB subscale, students are identified as at-risk if they score between 0 and 17 points and not at-risk if they score between 18 and 21.

Greater DIF effect sizes for males compared to females may be related to multiple factors. The larger effect sizes for males compared to females could be related to the student-teacher relationship, behavior expectations, implicit bias toward male students, or different behavior topography (Downey & Pribish, 2004; Hamre, Pianta, Downer, & Mashburn, 2008; Kunesh & Noltemeyer, 2019; O'Connor, Dearing, & Collins, 2010; Townsend, 2000). Nonetheless, DIF between different races and ethnicities and the interaction of race and ethnicity with gender has implications for placement decisions when using the SAEBRS EB subscale. Each student is identified as at-risk or not at-risk with the same cut scores.; however, the current study indicated that a student's score is influenced by group membership. The current study was not able to analyze sensitivity or specificity because there was no criterion measure; however, false positives and false negatives rates are likely influenced by DIF. For example, the average Black student is expected to score 1.68 raw score points lower (i.e., lower score is indicative of higher risk) compared to all other students on the EB subscale, when controlling for emotional behavior. Some Black students may be identified as at-risk due to DIF of the EB subscale. As a result, some researchers have suggested using multiple gating procedures, which may improve the quality for referrals (Severson et al., 2007).

Trends in DIF

Overall, positively worded items had the largest median effect and negatively worded items had smaller median effect sizes across racial/ethnic groups. Previous research has shown that teachers have more difficulty identifying internalizing behaviors in students (Herman et al., 2018). The current study indicated that although teachers may have had more difficulty identifying internalizing behaviors in their students, they were more consistent in applying the same criterion for negatively worded items (e.g., sadness and nervousness) across race and ethnicity, which contrasted with teachers' perceptions of the subjectivity of each item on the EB subscale of the SAEBRS. Teachers ranked negatively worded items as more subjective than positively worded items.

Explaining DIF

The IRQ questionnaire was created to explain the DIF effect sizes on the EB subscale of the SAEBRS. The null finding was likely due to the limited power of the micro-macro multilevel model in this study. However, the Kendall's tau correlation was significant. It was hypothesized that items that were perceived to be more subjective by teachers would display larger DIF by race/ethnicity. This hypothesis was not supported, and perceptions of item subjectivity was negatively correlated with the rank ordering of DIF. The items that teachers perceived as more subjective displayed smaller DIF and the items that teachers perceived as less subjective displayed larger DIF. It is unclear if this result is due to positively versus negatively worded items or some other underlying factor of the items. The DIF analyses indicated that teachers rated students differently on the EB subscale of the SAEBRS, and the Kendall's tau indicated that teachers' perception of item subjectivity did not align with ratings of students across race and ethnicity. The current study also supports the use of measurement invariance studies to examine item function across subgroups because perceptions of good versus bad items may not be supported through empirical analysis of the items.

Implications for Practice

The current study has implications for practice when using SEB universal screening tools. First, the current study revealed measurement invariance across race and ethnicity with the EB subscale of the SAEBRS, specifically for Black and White students. Practitioners should examine their screening data to determine if disproportionality across race and ethnicity in SEB risk identification is due to more problem behaviors or another factor associated with how the individual is rated on the measure. Problem-solving teams may evaluate the data through an ecological systems theory lens. At the individual level, teams may wish to evaluate individual student behaviors, how those behaviors are displayed in the context of the classroom/school, and any factors in the home or community that may be impacting disproportionate ratings. For example, behavior topography and behavioral expectations differ across races and ethnicities (Townsend, 2000). The problem-solving team should consider if cultural mismatches between students and teachers result in students are being rated differently because of behavior topography or behavior risk.

Schools and school districts may not be able to run the analyses in the current study, but they should consider calculating risk and risk ratios for racial and ethnic groups to determine if disproportionality exists with their own population. Local efforts may be warranted at the school or district level to increase rater accuracy (e.g., corrective feedback and teacher trainings) or consider using multiple gating procedures that may reduce false positive and false negative rates. Alternatively, districts or schools may wish to share data from risk and risk ratios with teachers in order to alert them of disproportionality that might exist in SEB ratings. Research has found that people that are internally motivated to act in a non-biased manner consider aspects of implicit bias after shown results of previously biased behavior (Fehr & Sassenberg, 2010; Fehr, Sassenberg, & Jonas, 2012). This method might be particularly beneficial for universal screening with BBRS in schools when screening is conducted multiple times per year. For example, a district using the SAEBRS for universal screening may indicate district-wide disproportionality in SEB risk by race and ethnicity in the fall. Prior to winter screening administration, participants can view fall data demonstrating disproportionate identification of minority students. According to the aforementioned hypothesis, teachers that are internally motivated to act with less bias may change their ratings of student behaviors. In this manner, no individual school or teacher is targeted as demonstrating biased ratings, given that district-wide data was shared. That is, at the student level implicit bias cannot be distinguished from actual behavioral problems. However, at the district or group level, it may be expected that behavior would be distributed closer to equal proportionality. In this manner, it may not be practical for districts to expect equal distribution of SEB risk across race and ethnicity. Instead, schools should be encouraged to monitor data at the classroom and school level to make informed decisions regarding disproportionate ratings of student behaviors.

Limitations

There are several limitations in the current study that need to be addressed and are organized by study. First, a unidimensional graded response model was used to examine DIF. However, the full SAEBRS uses a bifactor structure, which would indicate the use of a multidimensional model. Second, research has indicated that teachers are the greatest source of variability in universal BBRS (Tanner et al., 2018). Each teacher rated multiple student in their class. A multilevel model would control for the nested structure of responding, however the current study did not have access to teacher information. In addition, although the study calibration included over 11,000 students, it is possible that these students do not reflect national student demographics. For example, the sample demographics included a larger percentage of Black and White students, a smaller percentage of Hispanic students, and a smaller percentage of students in special education compared to the national average in 2016 (Snyder, de Brey, & Dillow, 2019). The smaller proportion of students in special education sampled may be due to their exclusion from universal screening practices, as many of these students may already have existing SEB data available for school use.

Study two assumed that discrimination and threshold parameters were stable after calibrating the items in study one. Therefore, a separate sample of teachers were surveyed on their perceptions of subjectivity of items. The independent sample of teachers may have differed in their perceptions from the teachers that completed the SAEBRS EB subscale was a third limitation. The current study attempted to control for this by sampling teachings that were already using the SAEBRS as part of their practice. Fourth, teachers may have different perceptions of subjectivity of items based off the student's race/ethnicity. The current study attempted to understand a total measurement invariance effect size across race and ethnicity. This may not have been the optimal analytic approach to understand DIF on the EB subscale of the SAEBRS. Student behaviors can be displayed differently across races and ethnicities (Townsend, 2000), which may impact how teachers rate and perceive the behaviors. Rather than computing a single overall effect size for each item, conducting analyses using the effect sizes for each students' race and ethnicity may provide more unique and beneficial data.

Fifth, information processing theory (IPT) did not significantly predict the absolute median effect sizes on the EB subscale. The current study utilized a micro-macro multilevel model (Croon & van Veldhoven, 2007). Simulation studies with this method have indicated that analyses required at least 100 groups to obtain adequate power (i.e., power = .80; Foster-Johnson & Kromrey, 2018). In the current study, the group size was limited by the number of items on the EB subscale of the SAEBRS (i.e., $n = 7$) because this was the number of median DIF effect sizes that were calculated. Future researchers may want to consider using measures that are longer. Alternatively, researchers may want to analyze the effect sizes for each race and ethnicity separately, rather than combined like in the current study. Analyzing the effect sizes separately for each race and ethnicity would create a larger sample size, therefore increasing power.

Lastly, the current study created a measure to evaluate the processes that raters consider when completing the EB subscale of the SAEBRS, using four steps as outlined by Tourangeau and Rasinski (1988). Other models of IPT include additional steps raters may consider when completing rating scales (Jobe, 2003), and such models may also provide different data (e.g., frequencies) than what was collected in the current study. Similar to how teachers may have different perceptions of subjectivity of items by race/ethnicity, teachers may go through different processes when rating items according to the race and/or ethnicity of the student.

Future Directions

There are several future directions for research that would extend the results of the current study. First, the current study was only conducted using the EB subscale of the SAEBRS. Future research could examine measurement invariance using the full SAEBRS scale. In addition, future studies could utilize a multilevel and multidimension IRT framework, in order to align with the bifactor structure of the SAEBRS. The current study attempted to understand why items were displaying DIF; however, additional explanatory IRT models could be used to better understand measurement invariance (De

Boeck & Wilson, 2004). In explanatory IRT models, the goal is to relate items on the test to variables related to the rater/examinee or aspects of the item. This model is useful when there are repeated measurement occasions or the testing situation is manipulated (De Boeck & Wilson, 2004). Future research should examine if aspects of items change DIF. For example, in the current study positively worded items displayed the greatest DIF. In future studies, items could be written with the opposite wording to determine if DIF was a function of positive versus negatively worded items (e.g., changing *'Difficulty rebounding from setbacks'* to '*Easily rebounds from setbacks'*). This may help provide greater clarity regarding when and how particular items may demonstrate DIF.

The current study used all other students not in the focal group for the reference group during DIF analyses (i.e., comparing White students to all other students in the sample). This was done so that DIF could be conducted on all racial/ethnic groups with large enough sample sizes and because White students should not set the criteria for comparison. In the current study, this meant that when Hispanic students and students with multiple races and ethnicities were the focal group, the reference group contained the negative response patterns for Black students and the positive response pattern for White students. This may have resulted in small DIF effect sizes for Hispanic students and students with multiple races and ethnicities. Future research may also wish to compare each racial/ethnic group to each other (e.g., White vs Black students or Hispanic vs. Asian students) in order to more fully understand how teachers' perceive and rate student behaviors between students from different race and/or ethnicity.

Lastly, future research should examine measurement invariance on a variety of other behavior rating scales to examine if results are consistent across measures.

Analyses could also be conducted on different scales with more questions to utilize the micro-macro multilevel model. Using scales with more questions (i.e., rating scales with more than 100 questions) would provide enough power for IPT to detect significance if it were to exist. However, this method may not be the most appropriate because of the number of questions that would need to be completed. For example, a 100-item scale would require teachers to rate four items from the IRQ on each of the 100 items, for a total of 400 responses.

Conclusion

This study provided additional evidence of the importance of conducting measurement invariance studies on BBRS. The results demonstrated teachers rate students differently across race and ethnicity, thus impacting the number of students identified as at-risk on the EB subscale of the SAEBRS. Specifically, larger effect sizes were identified with positively worded items compared to negatively worded items. In addition, measurement invariance effect sizes were larger for some groups compared to others. The preliminary findings indicate the need for further research attempting to explain measurement invariance with existing measures.

References

American Education Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing.* Washington, DC: American Education Research Association.

American Psychological Association Zero Tolerance Task Force. (2008). Are zero tolerance policies effective in the schools?: An evidentiary review and recommendations. *American Psychologist, 63,* 852-862. doi:10.1037/0003-066X.63.9.852

Arora, P. G., Connors, E. H., George, M. W., Lyon, A R., Wolk, C. B., & Weist, M. D. (2016). Advancing evidence-based assessment in school mental health: Key priorities for an applied research agenda. *Clinical Child and Family Psychology Review, 19,* 271-284. doi:10.1007/s10567-016-0217-y

Balu, R., Zhu, P., Doolittle, F., Schiller, E., Jenkins, J., & Gersten, R. (2015). Evaluation of response to intervention practices for elementary school reading (NCEE 2016-4000). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Banks, J. A. (2015). *Diversity and education.* New York, NY: Routledge.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173-1182.

Bates, D., Maechler, M., Bolker, B., & Walker, W. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1-48. doi:10.18637/jss.v067.i01

Benjamin, L. T. (2014). *A brief history of modern psychology* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Bennick, M., Croon, M. A., & Vermunt, J. (2013). Micro-macro multilevel analysis for discrete data: A latent variable approach and an application on personal network data. *Sociological Methods & Research, 42,* 431-457. doi:10.1177/0049124113500479

Benson, N. F., Floyd, R. G., Kranzler, J. H., Eckert, T. L., Fefer, S. A., & Morgan, G. B. (2019). Test use and assessment practice of school psychologists in the United States: Finding from the 2017 national survey. *Journal of School Psychology, 72,* 29-48. doi:10.1016/j.jsp.2018.12.004

Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice. 16,* 21-33. doi:10.1111/j.1745-3992.1997.tb00605.x

Boneshefski, M. J., & Runge, T. J. (2014). Addressing disproportionate discipline practices within a school-wide positive behavioral interventions and supports framework: A practical guide for calculating and using disproportionality rates. *Journal of Positive Behavior Interventions, 16,* 149-158. Doi:10.1177/1098300713484064

Bor, W., Dean, A. J., Najman, J. & Hayatbakhsh, R. (2014). Are child and adolescent mental health problems increasing in the 21st century?: A systematic review.

Australian & New Zealand Journal of Psychiatry, 48, 606-616. doi:10.1177/0004867414533834

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care, 44,* S176-S181. doi:10.1097/01.mlr.0000245143.08679.cc

Bottiani, J. H., Bradshaw, C. P., & Gregory, A. (2018). Nudging the gap: Introduction to the special issue "Closing in on Discipline Disproportionality". *School Psychology Review, 47,* 109-117. doi:10.17105/SPR-2018-0023.V47-2

Bowers, E. M. (1974). The primacy of primary prevention: The metaphor of screening. *School Psychology Review, 3,* 4-11.

Bradshaw, C. P., Koth, C. W., Thornton, L. A., & Leaf, P. J. (2009). Altering School Climate through school-wide positive behavioral interventions and supports: Findings from a group-randomized effectiveness trial. *Prevention Science, 10,* 100-115. doi:10.1007/s11121-008-0114-9

Bradshaw, C. P., Mitchell, M. M., & Leaf, P. J. (2010a). Examining the effects of schoolwide positive behavioral interventions and supports on student outcomes: Results from a randomized controlled effectiveness trial in elementary schools. *Journal of Positive Behavior Interventions, 12,* 133-148. doi:10.1177/1098300709334798

Bradshaw, C. P., Mitchell, M. M., O'Brennan, L. M., & Leaf, P. J. (2010b). Multilevel exploration of factors contributing to the overrepresentation of black students in office disciplinary referrals. *Journal of Educational Psychology, 102,* 508-520. doi:10.1037/a0018450

Brennan, L. M., Shaw, D. S., Dishion, T. J., & Wilson, M. N. (2015). The predictive utility of early childhood disruptive behaviors for school-age social functioning. *Journal of Abnormal Child Psychology, 43,* 1187-1199. doi:10.1007/s10802-014-9967-5

Brophy-Herb, H. E., Lee, R. E., Nievar, L. A., & Stollak, G. (2007). Preschoolers' social competence: Relations to family characteristics, teacher behaviors and classroom climate. *Journal of Applied Developmental Psychology, 28,* 134-148. doi:10.1016/j.appdev.2006.12.004

Brown, C. A., & Di Tillio, C. (2013). Discipline disproportionality among Hispanic and American Indian students: Expanding the discourse in U.S. research. *Journal of Education and Learning, 2,* 47-59.

Bruhn, A. L., Woods-Groves, S., & Huddle, S. (2014). A preliminary investigation of emotional and behavioral screening practices in K-12 schools. *Education and Treatment of Children, 37,* 611-634. doi:10.1353/etc.2014.0039

Burns, M. K., Appleton, J. J., & Stehouwer, J. D. (2005). Meta-analytic review of response to intervention research: Examining field-based and research-implemented models. *Journal of Psychoeducational Assessment, 23*, 381-394. doi:10.1177/073428290502300406

Cappelleri, J. C., Lundy, J. J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for quantitative assessment of items in developing patent-reported outcome measure, *Clinical Therapeutics, 36,* 648-662. doi:10.1016/j.clinthera.2014.04.006

Carroll, A., Houghton, S., Wood, R., Unsworth, K., Hattie, J., Gordon, L., & Bower, J. (2009). Self-efficacy and academic achievement in Australian high school students: The mediating effects of academic aspirations and delinquency. *Journal of Adolescence, 32,* 797-817.

Carter, A. S., Briggs-Gowan, M. J., & Ornstein Davis, N. (2004). Assessment of young children's social-emotional development and psychopathology: Recent advances and recommendations for practice. *Journal of Child Psychology & Psychiatry*, *45*, 109-134. doi:10.1046/j.0021-9630.2003.00316.x

Centers for Disease Control Foundation (n.d.). *What is public health?* Retrieved from http://www.cdcfoundation.org/content/what-public-health

Chalmers, P. R. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48,* 1-29. doi:10.18637/jss.v048.i06

Chang, D. F., & Demyan, A. (2007). Teachers' stereotypes of asian, black, and white students. *School Psychology Quarterly, 22,* 91-114. doi:10.1037/1045-3830.22.2.91

Chang, D. F., & Sue, S. (2003). The Effects of Race and Problem Type on Teachers' Assessments of Student Behavior. *Journal Consulting and Clinical Psychology, 71,* 235-241. doi:10.1037/0022-006X.71.2.235

Chatterji, P., Caffray, C. M., Crowe, M., Freeman, L., & Jensen, P. (2004). Cost assessment of a school-based mental health screening and treatment program in New York City. *Mental Health Service Research, 6,* 155-166. doi:10.1023/B:MHSR.0000036489.50470.cb

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of Statistical Software, 39,* 1-30.

Cohen, J. (1988). *Statistical power for the behavioral sciences.* (2nd ed.). Hillsdale, NJ: Earlbaum.

Cohen, J. (1992). A power primer, *Psychological Bulletin, 112,* 115-159.

Coie, J. D., Watt, N. F., West, S. G., Hawkins, J. D., Asarnow, J. R., Markman, H. J., ... Long, B. (1993). The science of prevention: A conceptual framework and some directions for a national research program. *American Psychologist*, *48*, 1013-1022. doi:10.1037/0003-066X.48.10.1013

Comrey, A., & Lee, H. (1992). *A first course in factor analysis*. Hillsdale, NJ: Erlbaum.

Cook, C. R., Frye, M., Slemrod, T., Lyon, A. R., Renshaw, T. L., & Zhang, Y. (2015). An integrated approach to universal prevention: Independent and combined effects of PBIS and SEL on youths' mental health. *School Psychology Quarterly, 30,* 166-183. doi:10.1037/spq0000102.

Cook, C. R., Volpe, R. J., & Livanis, A. (2010). Constructing a roadmap future universal screening practices beyond academics. *Assessment for Effective Intervention, 35,* 197-205. doi:10.1177/1534508410379842

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98-104

Crenshaw, K. (1993). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review, 43,* 1241-1299.

Croon, M. A., & van Veldhoven, M. J. P. M. (2007). Predicting group-level outcome variables from variables measured at the individual level: A latent variable multilevel model. *Psychological Methods, 12,* 45-57. doi:10.1037/1082-989X.12.1.45

Cullinan, D., & Epstein, M. H. (2013). *Emotional and behavioral screener*. Austin, TX: PRO•ED.

Davis, L. (1992). Instrument review: Getting the most from your panel of experts. *Applied Nursing Research, 5,* 194-197.

de Ayala, R. J. (2009). *The theory and practice of item response theory.* New York, NY: The Guilford Press.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach.* New York, NY: Springer.

De Los Reyes, A., Augenstein, T. M., Wang, M., Thomas, S. A., Drabick, D. A. G., Burgers, D. E., & Rabinowitz, J. (2015). The validity of the multi-informant approach to assessing child and adolescent mental health. *Psychological Bulletin, 141,* 858-900. doi:10.1037/a0038498

Dever, B. V., Dowdy, E., Raines, T. C., & Carnazzo, K. (2015). Stability and change of behavioral and emotional screening scores. *Psychology in the Schools, 52,* 818-629. doi:10.1002/pits.21825

Dever, B. V., Kamphaus, R. W., Dowdy, E., Raines, T. C., & DiStefano, C. (2013). Surveillance of middle and high school mental health risk by student self-report screener. *Western Journal of Emergency Medicine, 14,* 384-390. doi:10.5811/westjem.2013.2.15349

Dever, B. V., Raines, T. C., Dowdy, E., & Hostulter, C. (2016). Addressing disproportionality in special education using a universal screening approach. *The Journal of Negro Education, 85,* 59-71. doi:10.7709/jnegroeducation.85.1.0059

Dishion, T. J., & Kavanagh, K. (2000). A multilevel approach to family-centered prevention in schools: Process and outcome. *Addictive Behaviors, 25,* 899-911. doi:10.1016/S0306-4603(00)00126-X

DiStefano, C. A., & Kamphaus, R. W. (2007). Development and validation of a behavioral screener for preschool-age children. *Journal of Emotional & Behavioral Disorders, 15,* 93-102.

Domitrovich, C. E., Bradshaw, C. P., Greenberg, M. T., Embry, D., Poduska, J. M., & Ialongo, N. S. (2010). Integrated models of school-based prevention: Logic and theory. *Psychology in the Schools, 47,* 71-88. doi:10.1002/pits.20452

Dowdy, E., Dever, B. V., DiStefano, C., & Chin, J. K. (2011). Screening for emotional and behavioral risk among students with limited English proficiency. *School Psychology Quarterly, 26,* 14-26. doi:10.1037/a0022072

Dowdy, E., Doane, K., Eklund, K., & Dever, B. V. (2013). A comparison of teacher nomination and screening to identify behavioral and emotional risk within a sample of underrepresented students. *Journal of Emotional and Behavioral Disorders, 21,* 127-137. doi:10.1177/1063426611417627

Dowdy, E., Furlong, M., Raines, T. C., Bovery, B., Kauffman, B., Dever, B. V., … Murdock, J. (2015). Enhancing school-based mental health services with a preventive and promotive approach to universal screening for complete mental health. *Journal of Educational and Psychological Consultation, 25,* 175-197. doi:10.1080/10474412.2014.929951

Dowdy, E., Ritchey, K., & Kamphaus, R. W. (2010). School-Based Screening: A Population-Based Approach to Inform and Monitor Children's Mental Health Needs. *School Mental Health, 2*, 166-176. doi:10.1007/s12310-010-9036-3

Edelbrock, C., Costello, A. J., Dulcan, M. J., Conover, N. C., & Kala, R. (1986). Parent-child agreement on child psychiatric symptoms assessed via structured interview. *Journal of Child Psychology and Psychiatry, 27,* 181-190.

Eklund K., & Dowdy, E. (2014). Screening for behavioral and emotional risk versus traditional school identification methods. *School Mental Health, 6*, 40-49. doi: 10.1007/s12310-013-9109-1

Eklund, K., Renshaw, T. L., Dowdy, E., Jimerson, S. R., Hart, S. R., Jones, C. N., & Earhard, J. (2009). Early identification of behavioral and emotional problems in youth: Universal screening versus teacher-referral identification. *The California School Psychologist, 14,* 89-95.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

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

Farkas, G. (2003). Racial disparities and discrimination in education: What do we know, how do we know it, and what do we need to know? *Teachers College Record, 105*, 1119–1146. doi:10.1111/1467-9620.00279

Farmer, E. M. Z., Burns, B. J., Phillips, S. D., Angold, A., & Costello, E. J. (2003). Pathways into and through mental health services for children and adolescents. *Psychiatric Services, 54,* 60-66. doi:10.1176/appi.ps.54.1.60

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology, 50,* 229-238.

Fehr, J., & Sassenberg, K. (2010). Willing and able: How internal motivation and failure help to overcome prejudice. *Group Process & Intergroup Relations, 13,* 167-181. doi:10.1177/1368430209343116

Fehr, J., Sassenberg, K., & Jonas, K. J. (2012). Willful stereotype control: The impact of internal motivation to respond without prejudice on the regulation of activated stereotypes. *Zeitschrift für Psychologie, 220,* 180-186. doi:10.1027/2151-2604/a000111

Ferguson, R. F. (2003). Teachers' perceptions and expectations and the black-white test score gap. *Urban Education, 38,* 460-507. doi:10.1177/0042085903038004006

Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months, *Developmental Science, 16,* 234-248. doi:10.1111/desc.12019

Foster, S., Rollefson, M., Doksum, T., Noonan, D., Robinson, G., & Teich, J. (2005). *School Mental Health Services in the United States, 2002–2003. DHHS Pub. No. (SMA) 05-4068.* Rockville, MD: Center for Mental Health Services, Substance Abuse and Mental Health Services Administration.

Foster-Johnson, L., & Kromrey, J. D. (2018). Predicting group-level outcome variables: An empirical comparison of analysis strategies. *Behavioral Research Methods, 50,* 2461-2479. doi:10.3758/s13428-018-1025-8

Fuchs, D., & Fuchs, L. S. (2006). Introduction to response to intervention: What, why, and how valid is it? *Reading Research Quarterly, 4,* 93-99. doi:10.1598/RRQ.41.1.4

Fukuhara, H. & Kamata, A. (2011). A bifactor multidimensional item response theory model for differential item functioning analysis on testlet-based items. *Applied Psychological Measurement, 35,* 604-622. doi:10.1177/0146621611428447

Gage, N. A., Whitford, D. K., & Katsiyannis, A. (2018). A review of schoolwide positive behavior interventions and supports as a framework for reducing disciplinary exclusions. *The Journal of Special Education, 52*. 142-151. doi:10.1177/0022466918767847

Gerber, M. M., & Semmel, M. I. (1984). Teacher as imperfect test: Reconceptualizing the referral process. *Educational Psychologist, 16,* 137-148. doi:10.1177/0022219409335217

Gilliam, W. S., Maupin, A. N., Reyes, C. R., Accavitti, M., & Shic, F. (2016). *Do early educators' implicit biases regarding sex and race relate to behavior expectations and recommendations of preschool expulsions and suspensions?* Yale University Child Study Center. Retrieved from https://medicine.yale.edu/childstudy/zigler/publications/Preschool%20Implicit%20Bias%20Policy%20Brief_final_9_26_276766_5379_v1.pdf

Girvan, E. J., Deason, G., & Borgida, E. (2015, June 1). The generalizability of gender bias: Testing the effects of contextual, explicit, and implicit sexism on labor arbitration decisions. *Law and Human Behavior, 39,* 525-537. doi:10.1037/lhb0000139

Girvan, E. J., Gion, C., McIntosh, K., & Smolkowski, K. (2017). The relative contribution of subjective office referrals to racial disproportionality in school discipline. *School Psychology Quarterly, 32,* 392-404. doi:10.1037/spq0000178

Girio-Herrera, E., Dvorsky, M. R., & Owens, J. S. (2015). Mental Health screening in kindergarten youth: A multistudy examination of the concurrent and diagnostic validity of the Impairment Rating Scale. *Psychological Assessment, 27,* 215-227. doi:10.1037/a0037787

Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology, 45,* 117-135. doi:10.1016/j.jsp.2006.05.005

Gravois, T. A., & Rosenfield, S. A. (2006). Impact of instructional consultation teams on the disproportionate referral and placement of minority students in special education. *Remedial and Special Education, 27,* 42-52.

Green, J. G., Keenan, J. K., Guzmán, J., & Vinnes, S., Holt, M., & Comer, J. S. (2017). Teacher perspectives on indicators of adolescent social and emotional problems, *Evidence-Based Practice in Child and Adolescent Mental Health, 2,* 96-110. doi:10.1080/23794925.2017.1313099

Green, J. G., McLaughlin, K. A., Alergia, M., Costello, E. J., Gruber, M. J., Hoagwood, K., … Kessler, R. C. (2013). School mental health resources and adolescent mental health service use. *Journal of the American Academy of Child & Adolescent Psychiatry, 52,* 501-510. doi:10.1016/j.jaac.2013.03.002

Gupta, A., Szymanski, D. M., Leong, F. T. L. (2011). The "Model Minority Myth": Internalized racialism of positive stereotypes as correlates of psychological distress, and attitudes toward help-seeking. *Asian American Journal of Psychology, 2,* 101-114. doi:10.1037/a0024183

Hamre, B. K., Pianta, R. C., Downer, J. T., & Mashbum, A. J. (2008). Teachers' perceptions of conflict with young students: Looking beyond problem behaviors. *Social Development, 17*, 115-136. doi:10.1111/j.1467-9507.2007.00418.x

Harrison, J. R., Vannest, K. J., & Reynolds, C. R. (2013). Social acceptability of five screening instruments for social, emotional, and behavioral challenges. *Behavioral Disorders, 38,* 171-189. doi:10.1177/019874291303800305

Harry, B., & Anderson, M. G. (1994). The disproportionate placement of African American males in special education programs: A critique of the process. *Journal of Negro Education, 63,* 602-619.

Hartman, K., Gresham, F. M., & Byrd, S. (2017). Student internalizing and externalizing behavior screeners: Evidence for reliability, validity, and usability in elementary schools. *Behavioral Disorders, 42,* 108-118. doi:10.1177/0198742916688656

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin, 92,* 490-499.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analyses.* Orlando, FL: Academic Press, Inc.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analyses. *Psychological Methods, 3,* 486-504.

Herman, K. C., Cohen, D., Reinke, W. M., Ostrander, R., Burrell, L., McFarlane, E., & Duggan, A. K. (2018). Using latent profile and transition analyses to understand patterns of informant ratings of child depressive symptoms. *Journal of School Psychology, 69,* 84-99. doi:10.1016/j/jsp/2018.05.004

Hilhard, A. (1998). *SBA: Reawakening of the African mind*. Gainesville, FL: Makare.

Hinshaw, S. (1992). Externalizing behavior problems and academic underachievement in childhood and adolescence: Causal relationships and underlying mechanisms. *Psychological Bulletin, 111,* 127-155.

Holzinger, K., & Swineford, F. (1937). The Bi-factor method. *Psychometrika, 2*, 41-54. doi:10.1007/BF02287965

Horowitz, J., & Garber, J. (2006). The prevention of depressive symptoms in children and adolescents: A meta-analytic review. *Journal of Consulting and Clinical Psychology, 74,* 401-415. doi:10.1037/0022-006X.74.3.401

IDEA Data Center. (2014). *Methods for assessing racial/ethnic disproportionality in special education: A technical assistance guide* (Rev.). Rockville, MD: Westat, Julie Bollmer, Jim Bethel, Tom Munk, & Amy Bitterman.

Individuals with Disabilities Education Act, 20 U.S.C. § 1400 (2004).

Irvin, L. K., Tobin, T. J., Sprague, J. R., Sugai, G., & Vincent, C. G. (2004). Validity of office discipline referral measures as indices of school-wide behavioral status and effects of school-wide behavioral interventions. *Journal of Positive Behavior Interventions, 6,* 131-147. doi:10.1177/10983007040060030201

Jenkins, L. N., Demaray, M. K., Wren, N. S., Secord, S. M., Lyell, K. M., Magers, A. M., …Tennant, J. (2014). A critical review of five commonly used social-emotional and behavioral screeners for elementary and secondary schools. *Contemporary School Psychology, 18,* 241-254. doi:10.1007/s40688-014-0026-6

Jeon, M., Rijmen, F., Rabe-Hesketh, S. (2016). Modeling differential item functioning using a generalization of the multiple-group bifactor model. *Journal of Educational and Behavioral Statistics, 38,* 32-60. doi:10.3102/1076998611432173

Jimerson, S. R., Burns, M. K., & VanDerHeyden, A. M. (Eds.). (2015). *The handbook of response to intervention: Science and practice of multi-tiered systems of support* (2nd ed.). New York, NY: Springer Science.

Jobe, J. B., Tourangeau, R., & Smith, A. F. (1993). Contributions on survey research to the understanding of memory. *Applied Cognitive Psychology, 7,* 567-584.

Jobe, J. B. (2003). Cognitive psychology and self-reports: Models and methods. *Quality of Life Research, 21,* 219-227.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2,* 130-144. doi:10.1016/j.edurev.2007.05.002

Kamphaus, R. W. (2012). Screening for behavioral and emotional risk: Constructs and practicalities. *School Psychology Forum, 6,* 89-97.

Kamphaus, R. W., & Reynolds, C. R. (2007). *Behavior assessment system for children—second edition (BASC-2): Behavioral and Emotional Screening System (BESS).* Bloomington, MN: Pearson.

Kamphaus, R. W., & Reynolds, C. R. (2015). *Behavior assessment system for children, third edition (BASC-3) Behavioral and Emotional Student Screener.* San Antonio, TX: Pearson Education.

Kamphaus, R. W., Thorpe, J. S., Winsor, A. P., Kronke, A. P., Dowdy, E., & VanDeventer, M. C. (2007). Development and predictive validity of a teacher screener for child behavioral and emotional problems at school. *Psychological Measurement, 67,* 342-256. doi:10.1177001316440629204l

Kaufman, J. S., Vaughan, E. L., Reynolds, J. S., Donato, J. D., Bernard, S. N., & Hernandez-Brereton, M. (2010). Patters in office referral data by grade, race/ethnicity, and gender. *Journal of Positive Behavior Interventions, 12,* 44-54. doi:10.1177/1098300708329710

Kemper, A. R., Fant, K. E., Bruckman, D., & Clark, S. J. (2004). Hearing and vision screening program for school-aged children. *American Journal of Preventive Medicine, 26,* 141-146. doi:10.1016/j.amepre.2003.10.013

Kilgus, S. P., Chafouleas, S. M., & Riley-Tillman, T. C. (2013). Development and initial validation of the Social and Academic Behavior Risk Screener for elementary grades. *School Psychology Quarterly, 28*, 210-226. doi:10.1037/spq0000024

Kilgus, S. P., Chafouleas, S. M., Riley-Tillman, T. C., & von der Embse, N. P. (2014). *Social, Academic, and Emotional Behavior Risk Screener (SAEBRS).* Minneapolis, MN: Theodore J. Christ & Colleagues.

Kilgus, S. P., & Eklund, K. (2016). Consideration of base rates within universal screening for behavioral and emotional risk: A novel procedural framework. *School Psychology Forum, 10,* 120-130.

Kilgus, S. P., Eklund, K., von der Embse, N. P., Taylor, C., & Sims, W. A. (2016). Psychometric defensibility of the Social, Academic, and Emotional Behavior Risk Screener (SAEBRS) teacher rating scale and multiple gating procedure within elementary and middle school samples. *Journal of School Psychology, 58,* 21-39. doi:10.1016/j.jsp.2016.07.001

Kilgus, S. P., Sims, W., von der Embse, N. P., & Riley-Tillman, T. C. (2015). Confirmation of models for interpretation and use of the Social and Academic Behavior Risk Screener (SABRS). *School Psychology Quarterly, 30*, 335-352. doi:10.1037/spq0000087

Kilgus, S. P., Sims, W., von der Embse, N. P., & Taylor, C. N. (2015). Technical adequacy of the Social, Academic, and Emotional Behavior Risk Screener in an elementary sample. *Assessment for Effective Intervention, 42,* 46-59. doi:10.1177/1534508415623269

Kilgus, S. P., Taylor, C. N., & von der Emse, N. P. (2017). Screening for behavioral risk: Identification of high risk cut scores within the Social, Academic, and Emotional Behavior Risk Screener (SAEBRS). *School Psychology Quarterly, 33,* 155-159. doi:10.1037/spq0000230

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A Comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling, 18,* 212-228. doi:10.1080/10705511.2011.557337

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,* 39-116. doi:10.1111/j.1745-3984.2007.00029.x

King, K., Lembke, E., & Reinke, W. M. (2015). Using latent class analysis to identify academic and behavioral risk status in elementary students. *School Psychology Quarterly, 31,* 43-57. doi:10.1037/spq0000111

Kratochwill, T. R. (2007). Preparing psychologists for evidence-based school practice: Lessons learned and challenges ahead. *American Psychologist, 62,* 826-843. doi: 10.1037/0003-066X.62.8.829

Kristjansson, E., Aylesworth, R., McDowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement, 65,* 935-953. doi:10.1177/0013164405275668

Kunesh, C. E., & Noltemeyer, A. (2019). Understanding disciplinary disproportionality: Stereotypes shape pre-service teachers' beliefs about black boys' behavior. *Urban Education, 54,* 471-498. doi:10.1177/0042085915623337

Lambert, M. C., January, S. A., Cress, C. J., Epstein, M. H., & Cullinan, D. (2018). Differential item functioning across race and ethnicity for the Emotional and Behavioral Screener. *School Psychology Quarterly, 33,* 399-407. doi:10.1037/spq0000224

Lane, K. L., Carter, E. W., Jenkins, A., Dwiggins, L., & Germer, K. (2015). Supporting comprehensive, integrated, three-tiered models of prevention in schools: Administrators' perspectives. *Journal of Positive Behavior Interventions, 17,* 209-222. doi:10.1177/1098300715578916

Lau, A. S., Garland, A. F., Yeh, M., Mccabe, K. M., Wood, P. A., & Hough, R. L. (2004). Race/ethnicity and inter-informant agreement in assessing adolescent psychopathology. *Journal of Emotional and Behavioral Disorders, 12,* 145-156. doi:10.1177/10634266040120030201

Lee, W., Cho, S., McGugin, R. W., Van Gulick, A. B., & Gauthier, I. (2015). Differential item functioning analysis of the Vanderbilt Expertise Test for cars. *Journal of Vision, 15,* 1-19. doi:10.1167/15.13.23

Levitt, J. M., Saka, N., Romanelli, L. H., & Hoagwood, K. (2007). Early identification of mental health problems in schools: The status of instrumentation. *Journal of School Psychology, 45,* 163-191. doi:10.1016/j.jsp.2006.11.005

Lloyd, J. W., Kauffman, J. M., Landrum, T. J., & Roe, D. L. (1991). Why do teachers refer pupils for special education? An analysis of referral records. *Exceptionality, 2,* 115-126. doi:10.1080/09362839109524774

Lochman, J. E. (1995). Screening of child behavior problems for prevention programs at school entry. *Journal of Consulting and Clinical Psychology, 63,* 549-559. doi:10.1037/0022-006X.63.4.549

Loeber, R., Dishion, T. J., & Patterson, G. R. (1984). Multiple-gating: A multi-stage assessment procedure for identifying youths at-risk for delinquency. *Journal of Research on Crime and Delinquency, 21,* 7-32. doi:10.1177/0022427884021001002

Lorenzo, M. K., Frost, A. K., & Reinherz, H. Z. (2000). Social and emotional functioning of older Asian American adolescents. *Child and Adolescent Social Work Journal, 17,* 289-304.

Losen, D., Hodson, C., Keith, M. A., II, Morrison, K., & Belway, S. (2015). *Are we closing the school discipline gap?* Los Angeles, CA: The Center for Civil Rights Remedies at the Civil Rights Project of UCLA.

Lovett, J. M., Wolf, M., Frijters, J. C., & Steinbach, K. A. (2017). Early intervention for children at risk for reading disabilities: The impact of grade at intervention and individual differences on intervention outcomes. *Journal of Educational Psychology, 109,* 889-914. doi:10.1037/edu0000181

Low, S., Cook, C. R., Smolkowski, K., & Buntain-Ricklefs, J. (2015). Promoting social-emotional competence: An evaluation of the elementary version of Second Step®. *Journal of School Psychology, 53,* 463-477. doi:10.1016/j.jsp.2015.09.002

Lynn, M. (1986). Determination and quantification of content validity. *Nursing Research, 35,* 218-232.

MacMillan, D. L., Gresham, F. M., Lopez, M. F., & Bocia, K. M. (1996). Comparison of students nominated for prereferral interventions by ethnicity and gender. *The Journal of Special Education, 30,* 133-151.

Martella, R. C., Marchand-Martella, N. E., Woods, B., Thompson, S., Crockett, C., Northrup, E., … Ralston, N. C. (2010). Positive behavior support: Analysis of consistency between office discipline referrals and teacher recordings of disruptive classroom behaviors. *Behavioral Development Bulletin, 10,* 25-33. doi:10.1037/h0100517

Mason, B. A., Gunersel, A. B., & Ney, E. A. (2014). Cultural and ethnic bias in teacher ratings of behavior: A criterion-focused review. *Psychology in the Schools, 51,* 1014-1030. doi:10.1002/pits.21800

Mattison, R. E., Hooper, S. R., & Glassberg, L. A. (2002). Three-year course of learning disorders in special education students classified as behavioral disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, *41*, 1454-1461.

McCarthy, J. D., & Hoge D. R. (1987). The social construction of school punishment: Racial disadvantage out of universalistic process. *Social Forces, 65,* 1101-1120. doi:10.2307/2579025

McGrady, P. B., & Reynolds, J. R. (2013). Racial mismatch in the classroom: Beyond black-white difference. *Sociology of Education, 86,* 3-17. doi:10.1177/0038040712444857

McIntosh, K., Campbell, A. L., Carter, D. R., & Zumbo, B. D. (2009). Concurrent validity of office discipline referrals and cut points used in schoolwide positive behavior support. *Behavioral Disorders, 34,* 100-113. doi:10.1177/019874290903400204

McIntosh, K., Chard, D. J., Boland, J. B., & Horner, R. H. (2006). Demonstration of combined efforts in school-wide academic and behavioral systems and incidence of reading and behavior challenges in early elementary grades. *Journal of Positive Behavior Interventions, 8,* 146-154. doi:10.1177/10983007060080030301

McIntosh, K., Frank, J. L., & Spaulding, S. A. (2010). Establishing research-based trajectories of office discipline referrals for individual students. *School Psychology Review, 39,* 380-394.

McIntosh, K., Girvan, E. J., Horner, R. H., & Smolkowski, K. (2014). Education not incarceration: A conceptual model for reducing racial and ethnic disproportionality in school discipline. *Journal of Applied Research on Children: Informing Policy for Children at Risk, 5,* 1-22.

McKenzie, J. F., Wood, M. L., Kotecki, J. E., Clark, J. K., & Brey, R. A. (1999). Establishing content validity: Using qualitative and quantitative steps. *American Journal of Health Behaviors, 23,* 311-318. doi:10.5993/AJHB.23.4.9

Meade, A. W. (2010). A taxonomy of effect size measures in differential functioning of items and scales. *Journal of Applied Psychology, 95,* 728-743. doi:10.1037/a0018966

Mellard, D. F., McKnight, M., & Woods, K. (2009). Response to intervention screening and progress-monitoring practices in 41 local schools. *Learning Disabilities Research & Practice, 24,* 186-195. doi:10.1111/j.1540-5826.2009.00292.x

Mellard, D. F., Stern, A., & Woods, K. (2011). RTI school-based practices and evidence-based models. *Focus on Exceptional Children, 43,* 1-15.

Miller, F. G., Cohen, D., Chafouleas, S. M., Riley-Tillman, T. C., Welch, M. E., & Fabiano, G. A. (2015). A comparison of measures to screen social, emotional, and behavioral risk. *School Psychology Quarterly, 30,* 184-196. doi:10.1037/spq0000085

Morgan, P. L., Farkas, G., Tufis, P. A., & Sperling, R. A. (2008). Are reading and behavior problems risk factors for each other? *Journal of Learning Disabilities, 41,* 417-436.

National Association of School Psychologists. (2010). *Principles of professional ethics.* Bethseda, MD: Authors. Retrieved on July 3, 2018 from *https://www.nasponline.org/standards-and-certification/professional-ethics*

National Research Council. (2002). *Minority students in special and gifted education.* M. S. Donovan & C. T. Cross (Eds.). Washington, DC: National Academy Press.

Nguyen, T. H., Han, H., Kim, M. T., & Chan, K. S. (2014). An introduction to item response theory for patient-oriented outcome measurement. *Patient, 7,* 23-35. doi:10.1007/s40271-013-0041-0

O'Connor, E. E., Dearing, E., & Collins, B. A. (2011). Teacher–child relationship and behavior problem trajectories in elementary school. *American Educational Research Journal, 48,* 120-162. doi:10.3102/0002831210365008

Pearcy, M. T., Clopton, J. R., & Pope, A. W. (1993). Influences on teacher referral of children to mental health services: Gender, severity, and internalizing versus externalizing problems. *Journal of Emotional and Behavioral Disorders, 1,* 165-169. doi:10.1177/106342669300100304

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5, 50*, 157-175. doi:10.1080/14786440009463897

Pearson, A. R., Dovidio, J. F., & Gaertner, S. L. (2009). The nature of contemporary prejudice: Insights from aversive racism. *Social and Personality Psychology Compass, 3,* 314-338. doi:10.1111/j.1751-9004.2009.00183.x

Perepletchikova, F., Treat, T. A., & Kazdin, A. E. (2007). Treatment integrity in psychotherapy research: Analysis of studies and examination of the associated factors. *Journal of Consulting and Clinical Psychology, 75,* 829-841. doi:10.1037/0022-006X.75.6.829

Perou, R., Bitsko, R. H., Blumberg, S. J., Pastor, P., Ghandour, R. M., Gfroerer, J. C. … Huang, L. N. (2013). Mental Health Surveillance Among Children – United States, 2005–2011. *Morbidity and Mortality Weekly Report, 62,* 1-35.

Peshkin, A. (1988). In search of subjectivity—One's own. *Educational Researcher, 17,* 17-21. doi:10.3102/0013189X017007017

Predy, L., McIntosh, K., & Frank, J. L. (2014). Utility of number and type of office discipline referrals in predicting chronic problem behavior in middle schools. *School Psychology Review, 43,* 472-489.

Protection of Pupil Rights Amendment, 20 U.S.C. § 1232h (2002).

Puura, K., Almqvist, F., Tamminen, T., Piha, J., Kumpulainen, K., Rasanen, E., …& Koivisto, A. (1998). Children with depressed symptoms – What do the adults see? *Journal of Child Psychology & Psychiatry, 39,* 577-585. doi:10.1111/1469-7610.00353

Qi, C. H., & Kaiser, A. P. (2003). Behavior problems of preschool children from low-income families: Review of the literature. *Topic is Early Childhood Special Education, 23,* 188-216.

R Core Team. (2018). A language and environment for statistical computing (Version 3.5.2). Available from http://www.R-project.org/.

Raines, T. C., Dever, V. D., Kamphaus, R. W., & Roach, A. T. (2012). Universal Screening for Behavioral and Emotional Risk: A Promising Method for Reducing Disproportionate Placement in Special Education. *The Journal of Negro Education*, *81,* 283-296. doi:10.7709/jnegroeducation.81.3.0283

Randall, J., Cheong, Y. F., & Englehard, G. (2011). Using explanatory item response theory modeling to investigate context effects of differential item functioning for students with disabilities. *Educational and Psychological Measurement, 71,* 129-147. doi:10.1177/0013164410391577

Ready, D. D., & Wright, D. L. (2011). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities: The role of child background and classroom context. *American Educational Research Journal, 48,* 335-360. doi:10.3102/0002831210374874

Reeve, B. B., & Fayers, P. (2005). Applying item response theory modeling for evaluating questionnaire item and scale properties. *Assessing Quality of Life in Clinical Trials: Methods of Practice, 2,* 55-73.

Reise, S. P., Ainsworth, A. T., Haviland, M. G. (2005). Item response theory: Fundamentals, applications, and promises in psychological research. *Current Directions in Psychological Science, 14,* 95-101. doi:10.1111/j.0963-7214.2005.00342.x

Reise, S. P., Ventura, J., Nuechterlein, K. H., & Kim, K. H. (2005). An illustration of multilevel factor analysis. *Statistical Developments and Applications, 84,* 126-136. doi:10.1207/s15327752jpa8402_02

Reise, S. P., & Yu, J. (1990). Parameter Recovery in the Graded Response Model Using MULTILOG. *Journal of Educational Measurement, 27,* 133-144. doi:10.1111/j.1745-3984.1990.tb00738.x

Revicki, D. A., Chen, W. H., & Tucker, C. (2014). Developing item banks for patient reported health outcomes. In S. Reise & D. Revicki (Eds.), *Handbook of item* *response theory modeling: Applications to typical performance assessment* (pp. 334-363). New York: Routledge.

Reynolds, A. J., Ou, S., & Temple, J. A. (2018). A multicomponent, preschool to third grade preventive intervention and educational attainment at 35 years of age. *JAMA Pediatrics*, *172*, 247-256. doi:10.1001/jamapediatrics.2017.4673

Reynolds, C. R., & Kamphaus, R. W. (2015). *Behavior assessment system for children, third edition (BASC-3)*. San Antonio, TX: Pearson Education.

Romer, D., & McIntosh, M. (2005). The roles and perspectives of school mental health professionals in promoting adolescent mental health. In D. L. Evans, E. B. Foa, R. E. Gur, H. Hendin, C. P. O'Brien, M.E.P. Seligman, & B. T. Walsh (Eds.), *Treating and preventing adolescent mental health disorders: What we know and what we don't know* (pp. 598-615). New York: Oxford University Press.

Roque, M. (2010). Office discipline and student behavior: Does race matter? *American Journal of Education, 116,* 557-581. doi:10.1086/653629

Rubio, D. M., Berg-Weger, M., Tebb, S. S., Lee, E. S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study to social work research. *Social Work Research, 27,* 94-104. doi:10.1093/swr/27.2.94

Salvia, J., Ysseldyke, J., & Witmer, S. (2016). *Assessment: In special and inclusive education* (13^{th} ed.). Boston, MA: Cengage Learning.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, 34,* 1-97. doi:10.1007/BF03372160

Severson, H. H., Walker, H. M., Hope-Doolittle, J., Kratochwill, T. R., & Gresham, F. M. (2007). Proactive, early screening to detect behaviorally at-risk students: Issues, approaches, emerging innovations, and professional practices. *Journal of School Psychology, 45,* 193-223. doi:10.1016/j.jsp.2006.11.003

Schatschneider, C., Lane, K. L., Oakes, W. P., & Kalberg, J. R. (2014). The Student Risk Screening Scale: Exploring dimensionality and differential item functioning. *Educational Assessment, 19,* 185-203. doi:10.1080/10627197.2014.934608

Shockley, M. G. (2007). Literatures and definitions: Toward understanding Africentric education. *Journal of Negro Education, 76,* 103-117.

Shwartz, N. (1999). Self-reports: How questions shape answers. *American Psychologist, 54,* 93-105.

Skiba, R. J., Horner, R. H., Chung, C. G., Rausch, M. K., May, S. L., & Tobin, T. (2011). Race in not neutral: A national investigation of African American and Latino disproportionality in school discipline. *School Psychology Review, 40,* 85-107.

Skiba, R. J., Michael, R. S., Nardo, A. C., & Peterson, R. L. (2002). The color of discipline: Sources of racial and gender disproportionality in school punishment. *The Urban Review, 34,* 317-342. doi:10.1023/A:102132081

Skiba, R. J., Peterson, R. L., & Williams, T. (1997). Office referrals and suspension: Disciplinary intervention in middle schools. *Education and Treatment of Children, 20,* 295-315.

Smolkowski, K., Girvan, E. J., McIntosh, K., Nese, R. N. T., & Horner, R. H. (2016). Vulnerable decision points for disproportionate office discipline referrals: Comparisons of discipline for African American and White elementary school students. *Behavioral Disorders, 41,* 178-195. doi:10.17988/bedi-41-04-178-195.1

Snow, A. L., Cook, K. F., Lin, P., Morgan, R. O., & Magaziner, J. (2005). Proxies and other external raters: Methodological considerations. *Health Services Research, 40,* 1676-1693. doi:10.1111/j.1475-6773.2005.00447.x

Solomon, B. G., Klein, S. A., Hintze, J. M., Cressey, J. M., & Peller, S. L. (2012). A meta-analysis of school-wide positive behavior support: an exploratory study using single-case synthesis. *Psychology in the Schools, 49,* 105-121. doi:10.1002/pits.20625

Sprague, J. R. (2018). Closing in on discipline disproportionality: We need more theoretical, methodological, and procedural clarity. *School Psychology Review, 47,* 196-198. doi:10.17105/SPR-2018-0017.V47-2

Stice, E., Shaw, H., Bohon, C., Marti, C. N., & Rohde, P. (2009). A meta-analytic review of depression prevention programs for children and adolescents: Factors that predict magnitude of intervention effects. *Journal of Consulting and Clinical Psychology, 77,* 486-503. doi:10.1037/a0015168.

Stice, E., Shaw, H., & Marti, C. N. (2007). A meta-analytic review of obesity prevention programs for children and adolescents: The skinny on interventions that work. *Psychological Bulletin, 132,* 667-691.

Stone, A. A., Turkkan, J. S., Bachrach, C. A., Jobe, J. B., Kurtzman, H. S., & Cain, V. S. (2000). *The science of self-report.* Mahwah, NJ: Lawrence Erlbaum Associates

Sugai, G., & Horner, R. R. (2002). The evolution of discipline practices: School-wide positive behavior supports. *Child & Family Behavior Therapy, 24,* 23-50. doi:10.1300/J019v24n01_03

Sugai, G., & Horner, R. R. (2006). A promising approach for expanding and sustaining school-wide positive behavior support. *School Psychology Review, 35,* 245-259.

Sugai, G., Sprague, J. R., Horner, R. H., & Walker, H. M. (2000). Preventing school violence: The use of office discipline referrals to assess and monitor school-wide discipline interventions. *Journal of Emotional and Behavioral Disorders, 8,* 94-101. doi:10.1177/106342660000800205

Tanner, N., Eklund, K., Kilgus, S. P., & Johnson, A. H. (2018). Generalizability of universal screening measures for behavioral and emotional risk. *School Psychology Review, 47,* 3-17. doi:10.17105/SPR-2017-0044.V47-1

Tenenbaum, H. R., & Ruck, M. D. (2007). Are teachers' expectations different for racial minority than for European American students? A meta-analysis. *Journal of Educational Psychology, 99,* 253-273. doi:10.1037/0022-0663.99.2.253

Tilly, W. D. (2002). The evolution of school psychology to science-based practice: Problem solving and the three-tiered model. In A. Thomas & J. Grimes (Eds.), *best Practices in School Psychology: Volume 1* (pp. 17-36). Bethesda, MD: National Association of School Psychologists.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin, 103,* 299-314.

Townsend, B. L. (2000). The disproportionate discipline of African American learners: Reducing school suspensions and expulsions. *Exceptional Children, 66,* 381-391.

Turney, K., & McLanahan, S. (2015). The academic consequences of early childhood problem behaviors. *Social Science Research, 54,* 131-145. doi:10.1016/j.ssresearch.2015.06.022

Snyder, T.D., de Brey, C., & Dillow, S.A. (2019). *Digest of education statistics 2018* (NCES 2020-009). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

U.S. Department of Education, Office of Special Education and Rehabilitation Services. (2018). *39th Annual Report to Congress on the Implementation of the Individuals with Disabilities Education Act, 2017.* Washington, DC: Author.

U.S. Department of Health and Human Services, Centers for Medicare & Medicaid Services (2015). *Medicare and Medicaid milestones, 1937 to 2015.* Washington, DC: Author.

Vance, J. E., Bower, N. K., Fernandez, G., & Thompson, S. (2002). Risk and protective factors as predictors of outcome in adolescents with psychiatric disorder and aggression. *Journal of American Academy of Child and Adolescent Psychiatry, 41,* 36-43. doi:10.1097/00004583-200201000-00009

VanDerHayden, A. M., Burns, M. K., & Bonifay, W. (2018). Is more screening better? The relationship between frequent screening, accurate decisions, and reading proficiency. *School Psychology Review, 47,* 62-82. doi:10.17105/SPR-2017-0017.V47-1

VanDerHayden, A. M., Witt, J. C., & Gilbertson, D. (2007). A multi-year evaluation of the effects of a Response to Intervention (RTI) model on identification of children for special education. *Journal of School Psychology, 45,* 225-256. doi:10.1016/j.jsp.2006.11.004

Vincent, C. G., Tobin, T. J., Hawken, L. S., & Frank, J. L. (2012). Discipline referrals and access to secondary level support in elementary and middle schools: Patterns across African-American, Hispanic-American, and White students. *Education and Treatment of Children, 35,* 431-458.

von der Embse, N. P., Iaccarino, S., Mankin, A., Kilgus, S. P., Magen, E. (2017a). Development and validation of the Social, Academic, and Emotional Behavior Risk Screener-Student Rating Scale. *Assessment for Effective Intervention, 42,* 186-192. doi:10.1177/1534508416679410

von der Embse, N. P., Kilgus, S. P., Iaccarino, S., & Lei-Nielsen, S. (2017b). Screening for student mental health risk: Diagnostic accuracy, measurement invariance, and predictive validity of the Social, Academic, and Emotional Behavior Risk Screener-Student Rating Scale (SAEBRS-SRS). *School Mental Health, 9,* 273-283. doi:10.1007/s12310-017-9214-7

Wagner, M., Kutash, K., Duchnowski, A. J., Epstein, M. H., & Sumi, W. C. (2005). The children and youth we serve: A national picture of the characteristics of students with emotional disturbances receiving special education. *Journal of Emotional and Behavioral Disorders, 13,* 79-96. doi:10.1177/10634266050130020201

Walker, H. M., Nishioka, V. M., Zeller, R., Severson, H. H., & Feil, E. G. (2000). Causal factors and potential solutions for the persistent under identification of students having emotional or behavioral disorders in the context of schooling. *Assessment for Effective Intervention, 26,* 29-39. doi:10.1177/073724770002600105

Walker, H. M., & Severson, H. H. (2014). *Systematic Screening for Behavioral Disorders* (2nd ed.). Eugene, OR: Pacific Northwest Publishing.

Wallace, J. M., Goodkind, S., Wallace, C. M., & Bachman, J. G. (2008). Racial, ethnic, and gender differences in school discipline among U.S. high school students: 1991-2005. *The Negro Educational Review, 59,* 47–62.

Weber, T. (2003). There is no objective subjectivity in the study of social interaction. *Forum Qualitative Social Research, 4,* 1-18. doi:10.17169/fqs-4.2.716

Webster-Stratton, C., & Reid, M. J. (2003). Treating conduct problems and strengthening social and emotional competence in young children: The Dina Dinosaur Treatment Program. *Journal of Emotional and Behavioral Disorders, 11,* 130-14.

Weisz, J. R., Weisz, B., Han, S. S., Granger, D. A., & Morton, T. (1995). Effects of psychotherapy with children and adolescents revisited: A meta-analysis of treatment outcome studies. *Psychological Bulletin, 117,* 450-468.

Woltman, H. Feldstain, A., MacKay, J. C., & Rocchi, M. (2012). An introduction to hierarchical linear modeling. *Tutorials in Quantitative Methods for Psychology, 8,* 52-69. doi:10.20982/tqmp.08.1.p052

Wood, P. A., Yeh, M., Pan, D., Lambros, K. M., McCabe, K. M., & Hough, R. L. (2005). Exploring the relationship between race/ethnicity, age of first school-based services utilization, and age of first specialty mental health care for at-risk youth. *Mental Health Services Research, 7,* 185-196. doi:10.1007/s11020-005-5787-0

Woodman, A. C., Demers, L., Crossman, M. K., Warfield, M. E., & Hauser-Cram, P. (2018). Part C Early Intervention dosage and growth in adaptive skills from early childhood through adolescence. *Early Childhood Research Quarterly, 43,* 73-82. doi:10.1016/j.ecresq.2018.01.007

Whitford, D. K., & Whitford & Levine-Donnerstein, D. (2014). Office disciplinary referral patterns of American Indian students from elementary school through high school. *Behavioral Disorders, 39,* 78-88.

Wright, J. A., & Dusek, J. B. (1998). Compiling school base rates for disruptive behaviors from student disciplinary referral data. *School Psychology Review, 27,* 138-147.

Young, E. L., Sabbah, H. Y., Young, J. B., Resier, M. L., & Richarson, M. J. (2010). Gender differences and similarities in a screening process for emotional and behavioral risks in secondary schools. *Journal of Emotional and Behavioral Disorders, 18,* 225-235. doi:10.1177/1063426609338858

Zhang, D., Katsiyannis, A., Ju, S., & Roberts, E. (2014). Minority representation in special education: A 5-year trend. *Journal of Child and Family Studies, 23*, 118-127. doi:10.1007s10826-012-9698-6

Zimmermann, C. R. (2018). The penalty of being a young black girl: Kindergarten teachers' perceptions of children's problem behaviors and student-teacher conflict by the intersection of race and gender. *The Journal of Negro Education, 87,* 154-168. doi:10.7709/jnegroeducation.87.2.0154

Zuckerbrot, R. A., Cheung, A., Jensen, P. S., Stein, R. E. K., Laraque, D., & GLAD-PC STEERING GROUP. (2018). Guidelines for adolescent depression in primary care (GLAD-PC): Part I. Practice preparation, identification, assessment, and initial management. *Pediatrics, 141,* 1-21. doi:10.1542/peds.2017-4081

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic Regression modeling as a unitary framework for* *binary and likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4,* 223-233.

Appendix A

Table A1

Discrimination and threshold parameters for Asian students (n = 291) on the Emotional Behavior Subscale of the Social, Academic, and Emotional Behavior Risk Screener.

		Difficulty Thresholds		
Item	Discrimination	*0 – 1*	*1 – 2*	*2 – 3*
Adaptability	1.93	-2.63	-1.60	0.49
	(.29)	(.31)	(.17)	(.10)
Difficulty rebounding from setbacks	2.29	-2.57	-2.03	-0.98
	(.41)	(.30)	(.21)	(.11)
Nervousness	1.73	-4.00	-3.17	-1.47
	(.35)	(.82)	(.52)	(.19)
Positive attitude	2.60	-3.25	-1.62	-0.63
	(.43)	(.50)	(.16)	(.10)
Sadness	3.14	-3.10	-2.51	-1.08
	(.65)	(.45)	(.29)	(.11)
Withdrawal	1.93	-3.41	-2.57	-0.96
	(.31)	(.52)	(.31)	(.13)
Worry	1.65	-3.72	-2.67	-0.61
	(.27)	(.62)	(.35)	(.12)

Note. Standard errors are in parentheses.

Table 2A

Discrimination and threshold parameters for Native American students (n = 87) on the Emotional Behavior Subscale of the Social, Academic, and Emotional Behavior Risk Screener.

		Difficulty Thresholds		
Item	Discrimination	*0 – 1*	*1 – 2*	*2 – 3*
Adaptability	1.18	-3.28	-0.85	0.62
	(.31)	(.88)	(.28)	(.27)
Difficulty rebounding from setbacks	1.43	-2.35	-1.82	-0.08
	(.35)	(.53)	(.40)	(.21)
Nervousness	5.34	-2.02	-1.45	-0.49
	(1.95)	(.34)	(.22)	(.14)
Positive attitude	1.20	-4.22	-1.26	0.41
	(.33)	(1.26)	(.35)	(.25)
Sadness	3.49	-2.19	-1.66	-0.08
	(.92)	(.38)	(.27)	(.15)
Withdrawal	1.90	-2.45	-1.43	-0.19
	(.44)	(.51)	(.28)	(.18)
Worry	5.82	-2.00	-1.32	-0.18
	(2.37)	(.33)	(.20)	(.14)

Note. Standard errors are in parentheses.

Table 3A

Discrimination and threshold parameters for male students with multiple races/ethnicities *(n = 324) on the Emotional Behavior Subscale of the Social, Academic, and Emotional Behavior Risk Screener.*

		Difficulty Thresholds		
Item	Discrimination	*0 – 1*	*1 – 2*	*2 – 3*
Adaptability	1.61	-2.48	-1.13	0.11
	(.20)	(.28)	(.14)	(.10)
Difficulty rebounding from setbacks	2.06	-1.97	-1.44	-0.28
	(.24)	(.19)	(.14)	(.09)
Nervousness	3.13	-3.10	-2.31	-1.04
	(.50)	(.43)	(.23)	(.10)
Positive attitude	2.16	-2.71	-1.13	0.21
	(.26)	(.29)	(.12)	(.09)
Sadness	5.14	-2.18	-1.59	-0.62
	(.97)	(.19)	(.12)	(.08)
Withdrawal	2.48	-2.33	-1.68	-0.57
	(.31)	(.22)	(.15)	(.09)
Worry	2.48	-2.86	-1.81	-0.52
	(.33)	(.33)	(.17)	(.09)

Note. Standard errors are in parentheses.

Table 4A

Discrimination and threshold parameters for female students with multiple races / ethnicities (n = 318) on the Emotional Behavior Subscale of the Social, Academic, and Emotional Behavior Risk Screener.

		Difficulty Thresholds		
Item	Discrimination	*0 – 1*	*1 – 2*	*2 – 3*
Adaptability	2.37	-2.49	-1.14	-0.25
	(.37)	(.27)	(.12)	(.10)
Difficulty rebounding from setbacks	1.91	-2.32	-1.89	-0.63
	(.27)	(.25)	(.19)	(.10)
Nervousness	1.77	NA[a]	-3.18	-1.57
	(.35)	NA[a]	(.51)	(.20)
Positive attitude	2.73	-2.89	-1.37	-0.33
	(.45)	(.37)	(.13)	(.09)
Sadness	2.48	-2.71	-2.09	-0.70
	(.45)	(.33)	(.22)	(.10)
Withdrawal	2.59	-2.99	-1.89	-0.82
	(.42)	(.39)	(.18)	(.10)
Worry	1.56	-3.92	-2.69	-0.78
	(.26)	(.66)	(.36)	(.13)

Note. Standard errors are in parentheses. [a]No responses of *Never* for this group

Appendix B

Initial Evaluation of the Item Response Questionnaire

Instructions: The purpose of the proposed measure is to gain an understanding of the factors teachers consider when completing rating scales of student behavior. This measure will be completed after completion of a student behavior rating scale. The goal of this project is to provide researchers and practitioners data to better understand the defensibility of rating scale data, as well as any decisions that might result from that data. Please read each of the items carefully and answer three ratings for each of the items on the behavior rating scale.

Below is a four-step process that respondents go through to complete ratings of their attitudes and perceptions (Tourangeau & Rasinski, 1988):

Step 1: Comprehension of the question	How the individual understands and interprets the behavior they are being asked to rate.
Step 2: Memory	The individual's ability to recall events from his/her memory when rating a student's behavior.
Step 3: Decision Making	How the individual arrives at an estimate of the frequency of a behavior based off the events recalled from memory.
Step 4: Formulation of Response	How the individual places their estimate of the frequency of a behavior onto the provided response options and checks that the response is consistent with their previous responses.

1) Indicate which question best represents the step by ranking their order (with 1 being the best representation of the step).

2) Indicate how confident you are with your choice.

NC = Not Confident	SC = Somewhat Confident	MC = Mostly Confident	VC = Very Confident

3) Indicate how relevant the statement is to the step.

NR = Not Relevant	SR = Somewhat Relevant	MR = Mostly Relevant	VR = Very Relevant

Example:

Question	Rank	Confidence	Relevance
Question 1	3	NC SC MC VC	NR SR MR VR
Question 2	1	NC SC MC VC	NR SR MR VR
Question 3	2	NC SC MC VC	NR SR MR VR

Thank you very much for you time!

Teachers will complete the items below while considering each individual item from a behavior rating scale. Therefore, the respondent will complete these same questions multiple times when considering a single rating scale.

Step	Question	Rank	Confidence	Relevance
1	I understand this question.		NC SC MC VC	NR SR MR VR
	I can recognize this behavior when it is displayed.		NC SC MC VC	NR SR MR VR
	My definition of this behavior is similar to other people's definition of this behavior.		NC SC MC VC	NR SR MR VR
	I have seen my students display this behavior before.		NC SC MC VC	NR SR MR VR
2	I use all events of the behavior when rating this behavior.		NC SC MC VC	NR SR MR VR
	I can remember behaviors related to this question.		NC SC MC VC	NR SR MR VR
	I use specific/discrete events of this behavior to rate this question.		NC SC MC VC	NR SR MR VR
3	I weigh each instance of the student's behavior I recall equally when responding to this question.		NC SC MC VC	NR SR MR VR
	I rate the frequency of this behavior based off an estimation of how often the student engages in the behavior.		NC SC MC VC	NR SR MR VR
	I consider the intensity of the student's behavior when responding to this question.		NC SC MC VC	NR SR MR VR
4	I would rate this behavior similarly to other teachers.		NC SC MC VC	NR SR MR VR
	The response options (e.g., Never, Sometimes, Often, and Almost Always) represent the frequency of this behavior.		NC SC MC VC	NR SR MR VR
	My ratings of this behavior would be similar to other people's ratings.		NC SC MC VC	NR SR MR VR
	I compare my previous ratings on other behaviors when rating this question.		NC SC MC VC	NR SR MR VR

Are there any questions missing from any category? If so, please write the question and category.

Appendix C

Table 1C

Final four questions for each of the four steps in Information Processing Theory.

Question
1. I can recognize this behavior when it is displayed.
2. I use specific/discrete events of this behavior when rating this question.
3. I rate the frequency of this behavior based off an estimation of how often the student engages in the behavior.
4. I compare my previous ratings on other behaviors when rating this question

Appendix D

Figure D1

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for Asian students on item 1.

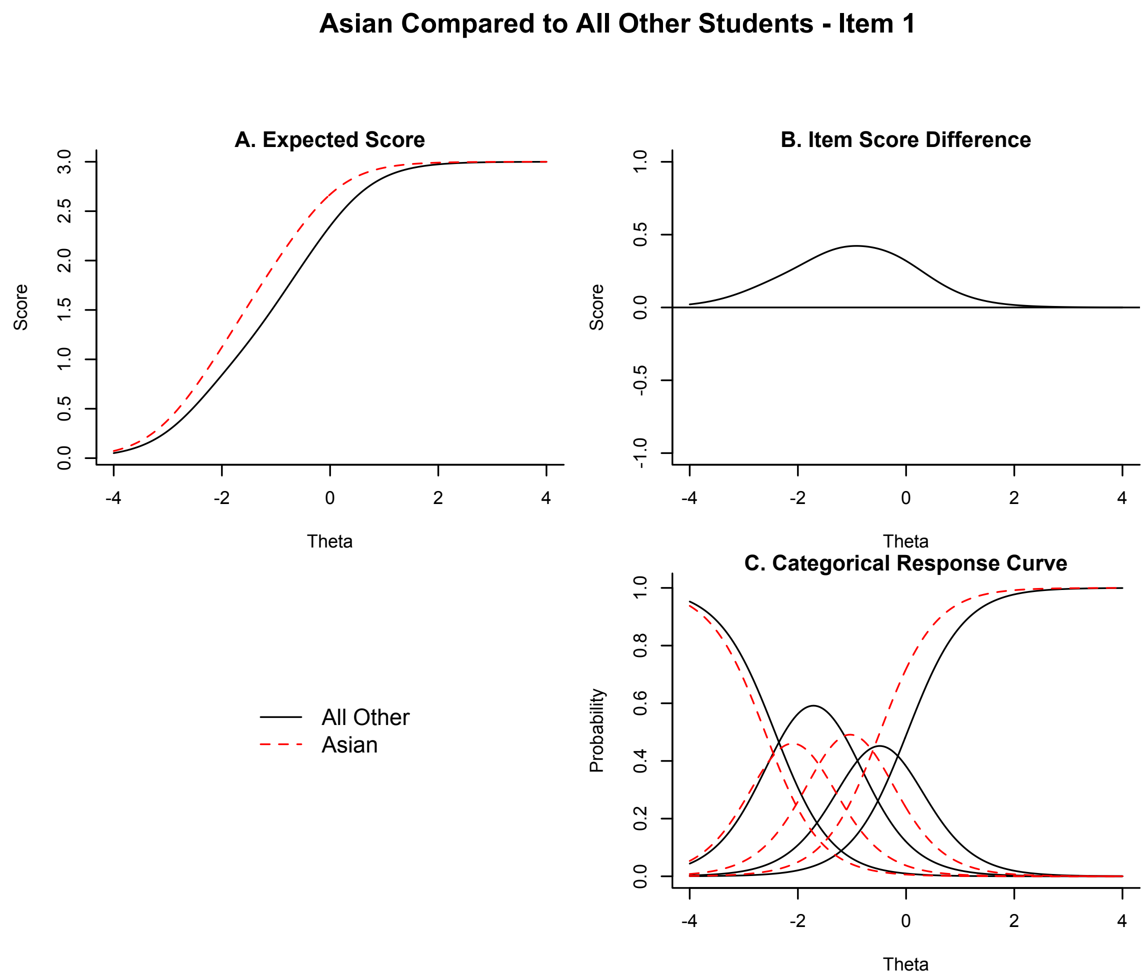


Figure D2

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for Asian students on item 2.

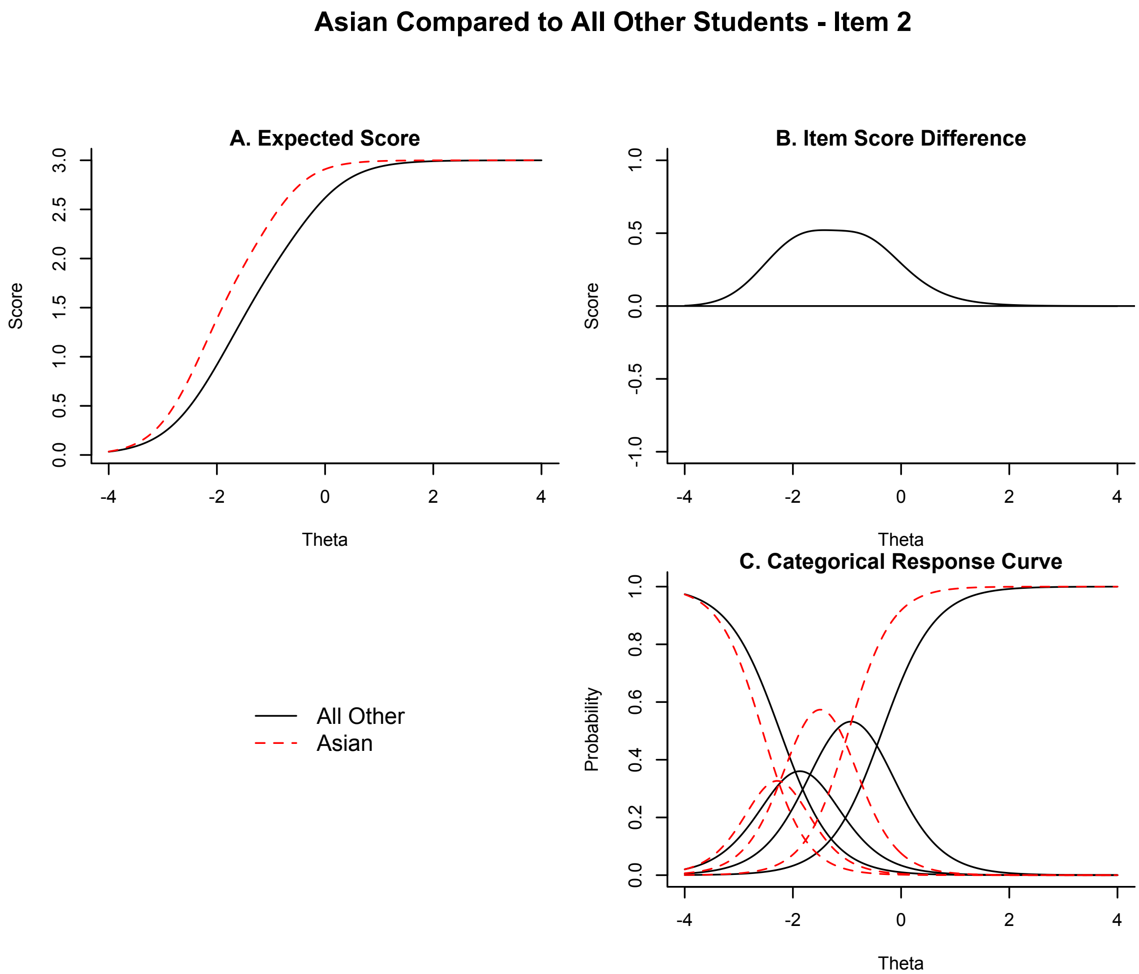


Figure D3

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for Asian students on item 3.

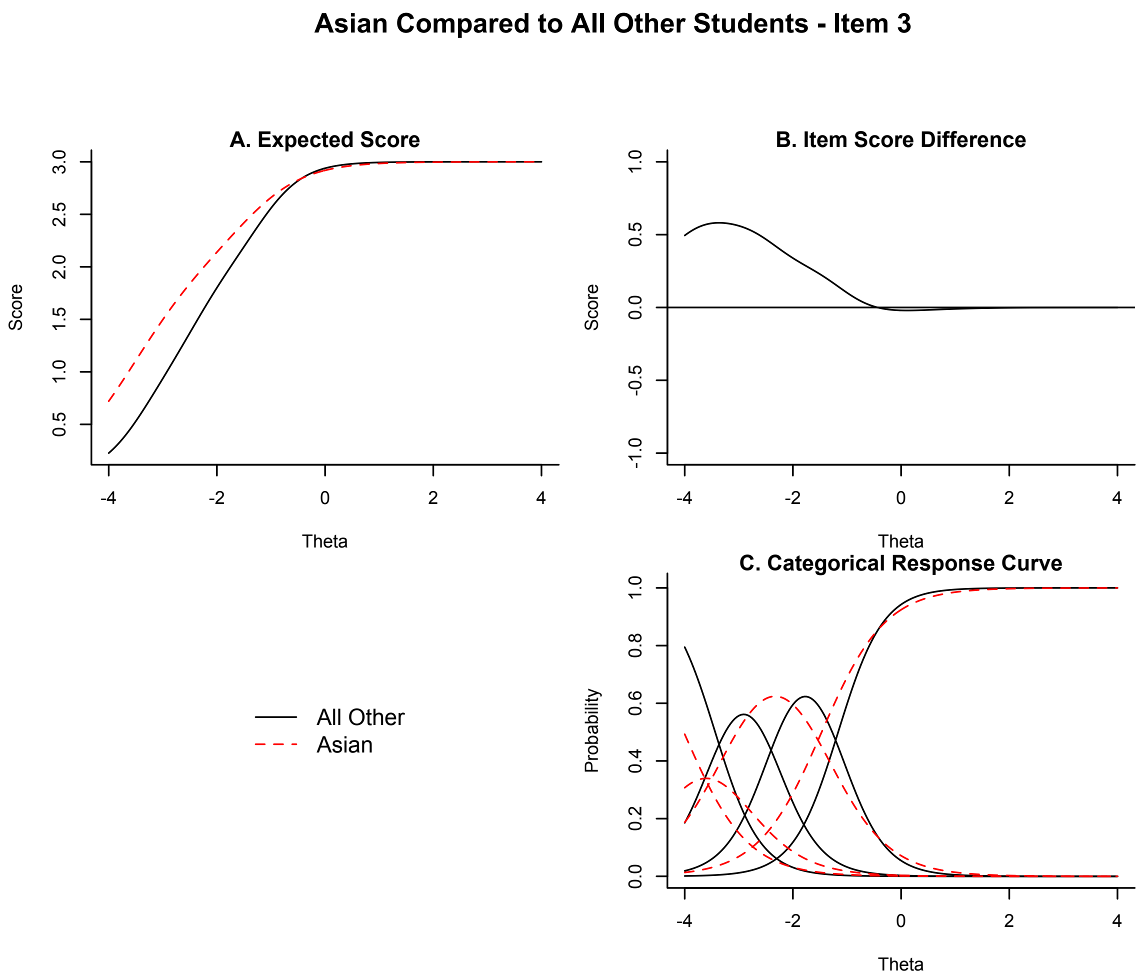


Figure D4

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for Asian students on item 4.

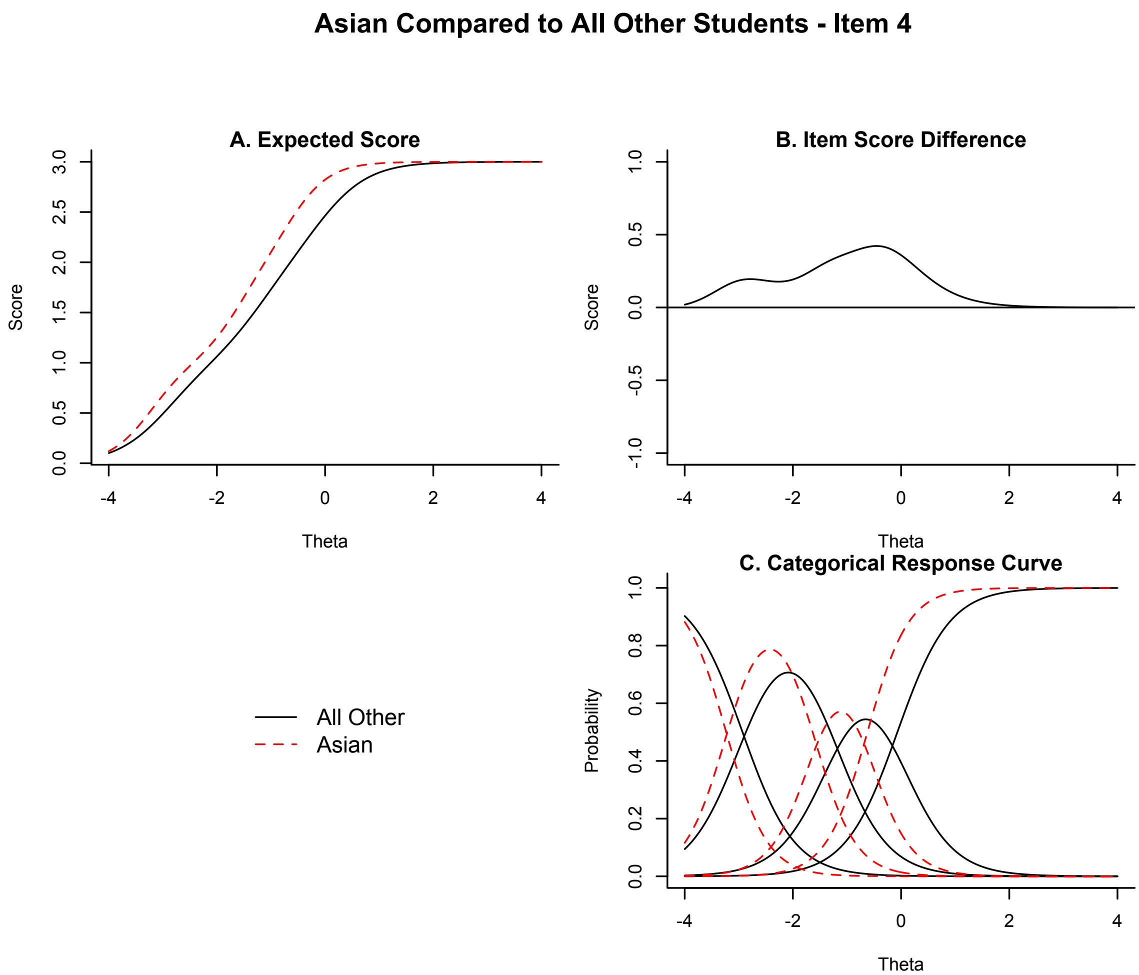


Figure D5

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for Asian students on item 5.

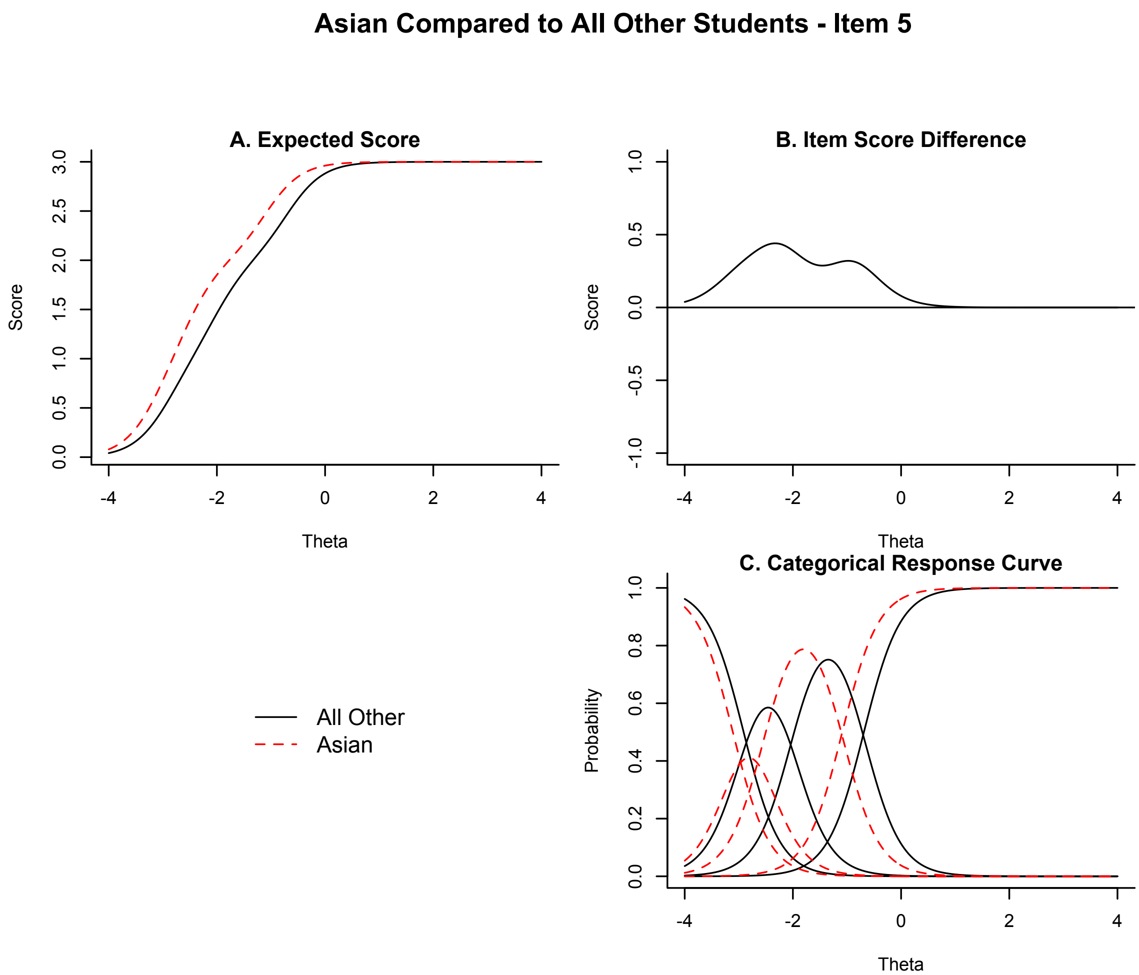


Figure D6

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for Asian students on item 6.

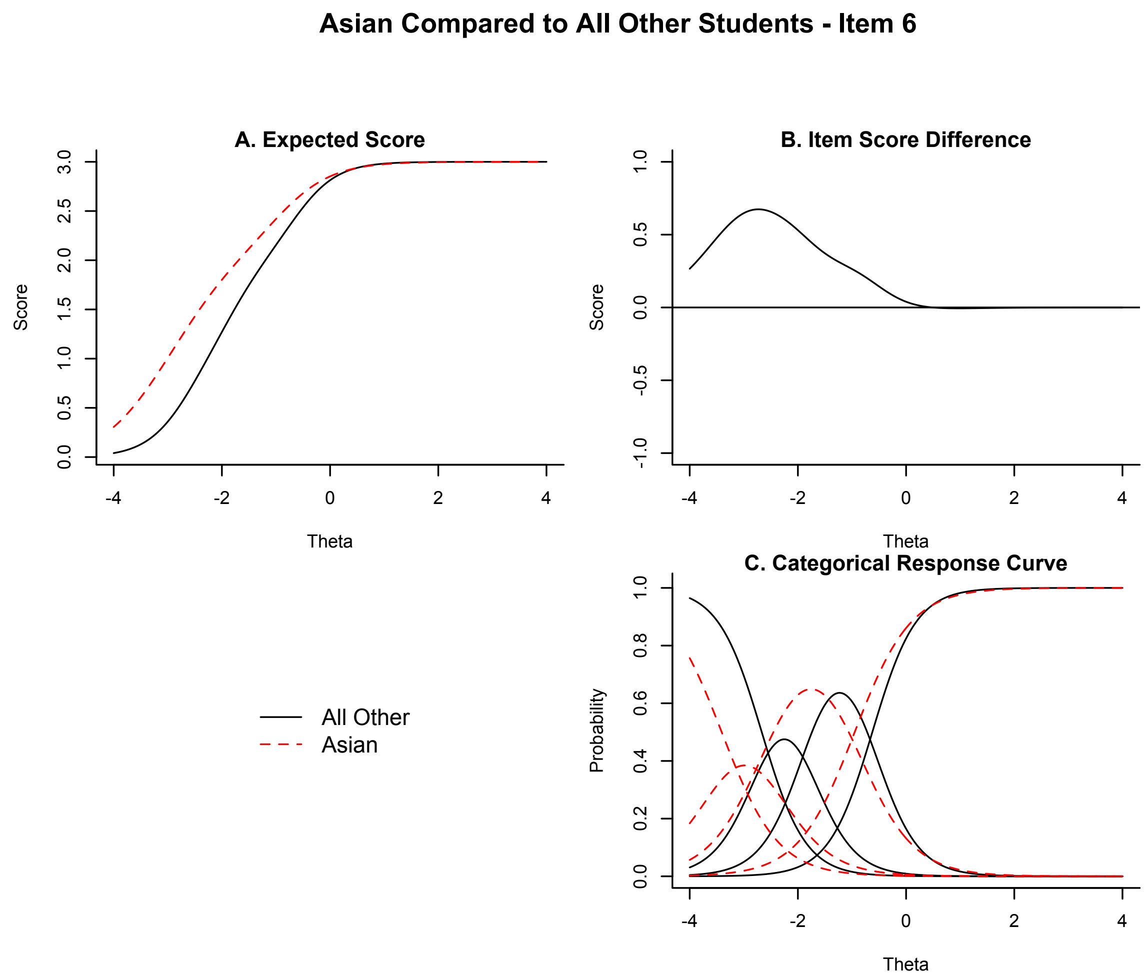


Figure D7

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for Asian students on item 7.

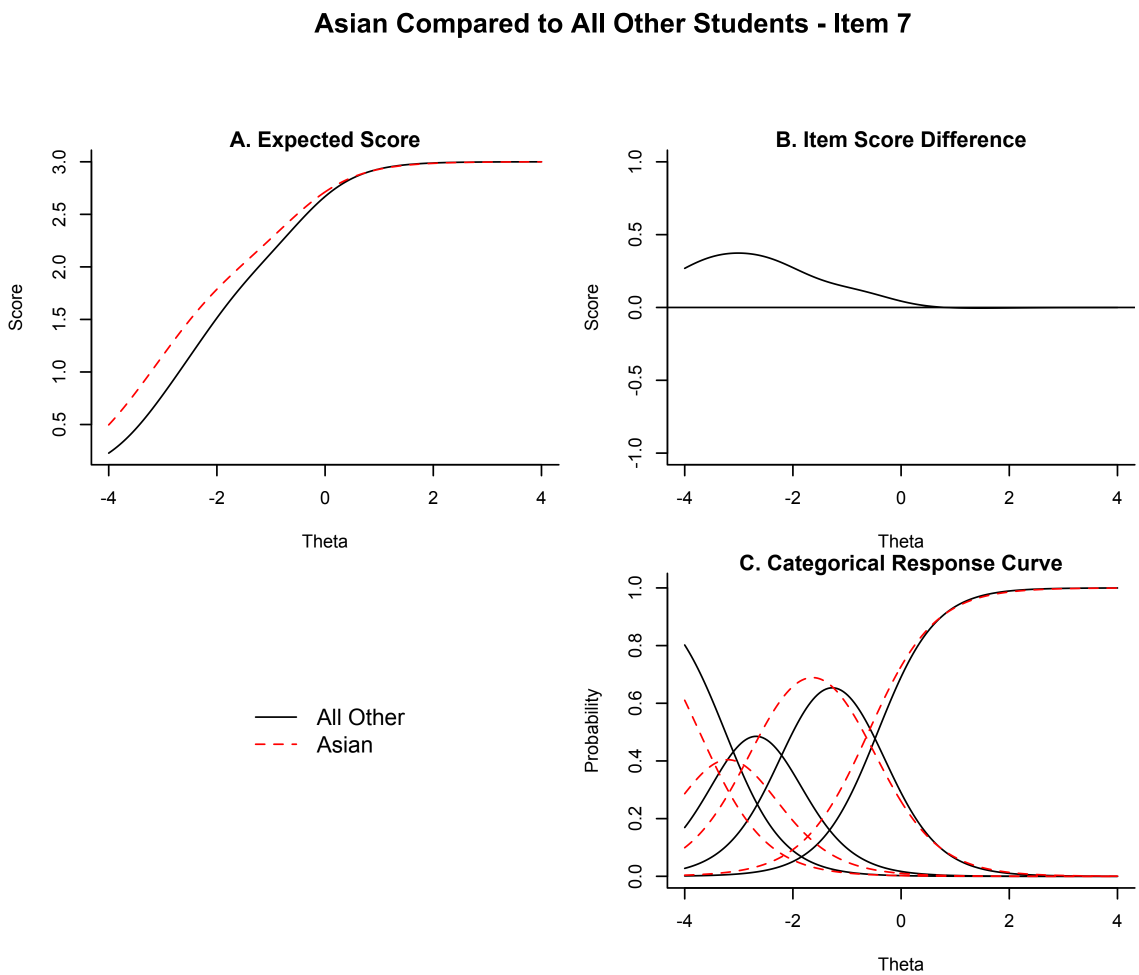


Figure D8

Graphs displaying the test response functions and the difference between the item response functions for Asian students.

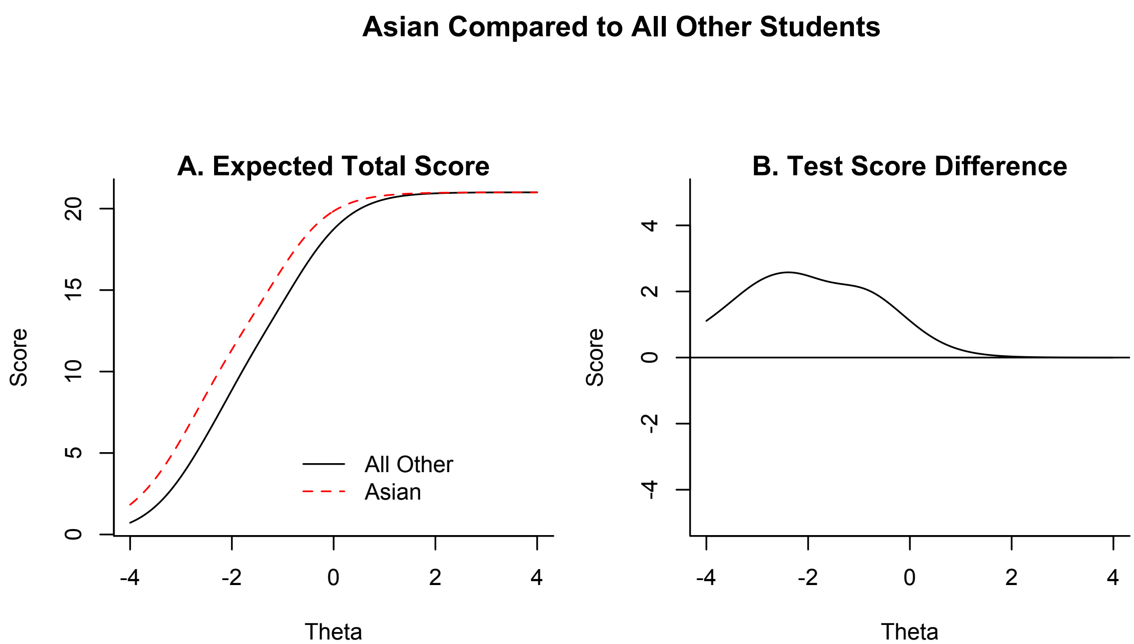


Figure D9

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for Black students on item 1.

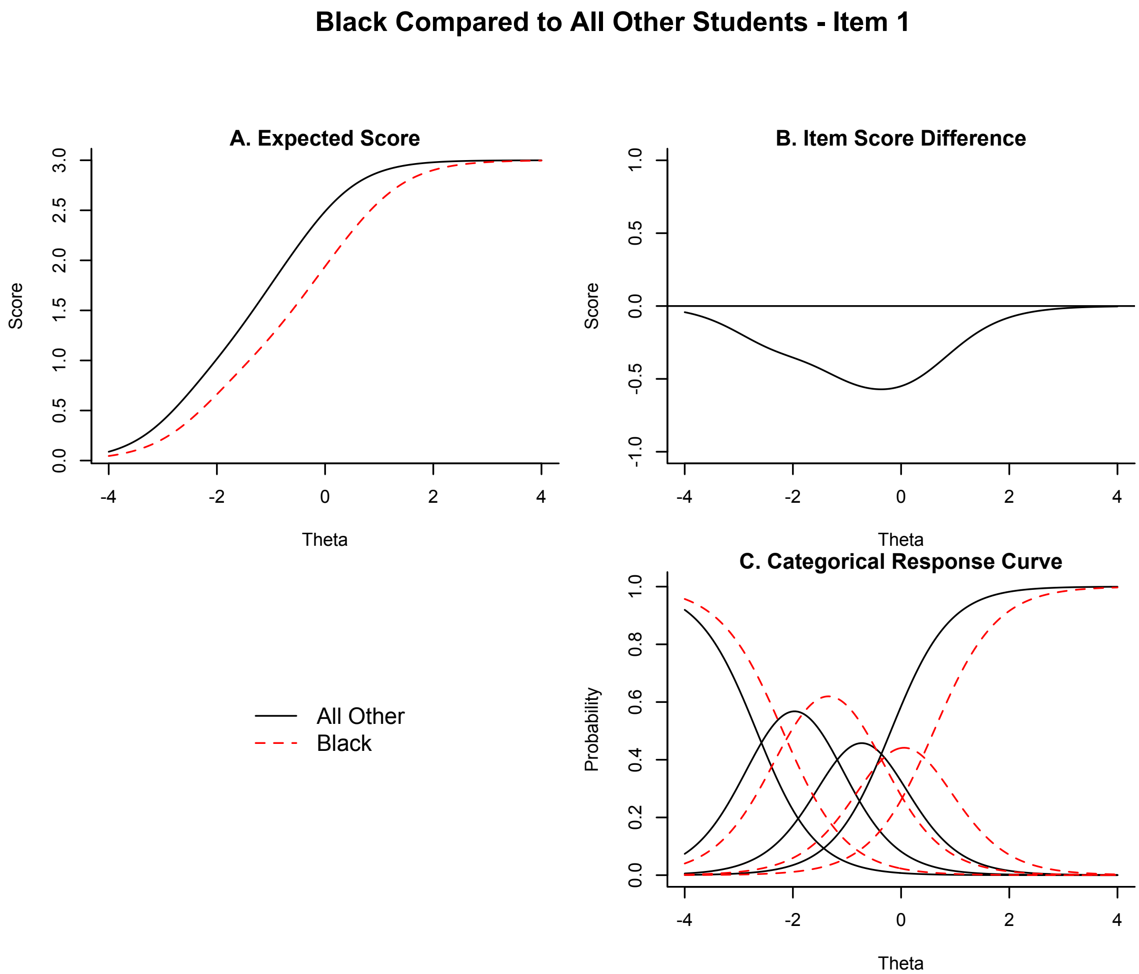


Figure D10

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for Black students on item 2.

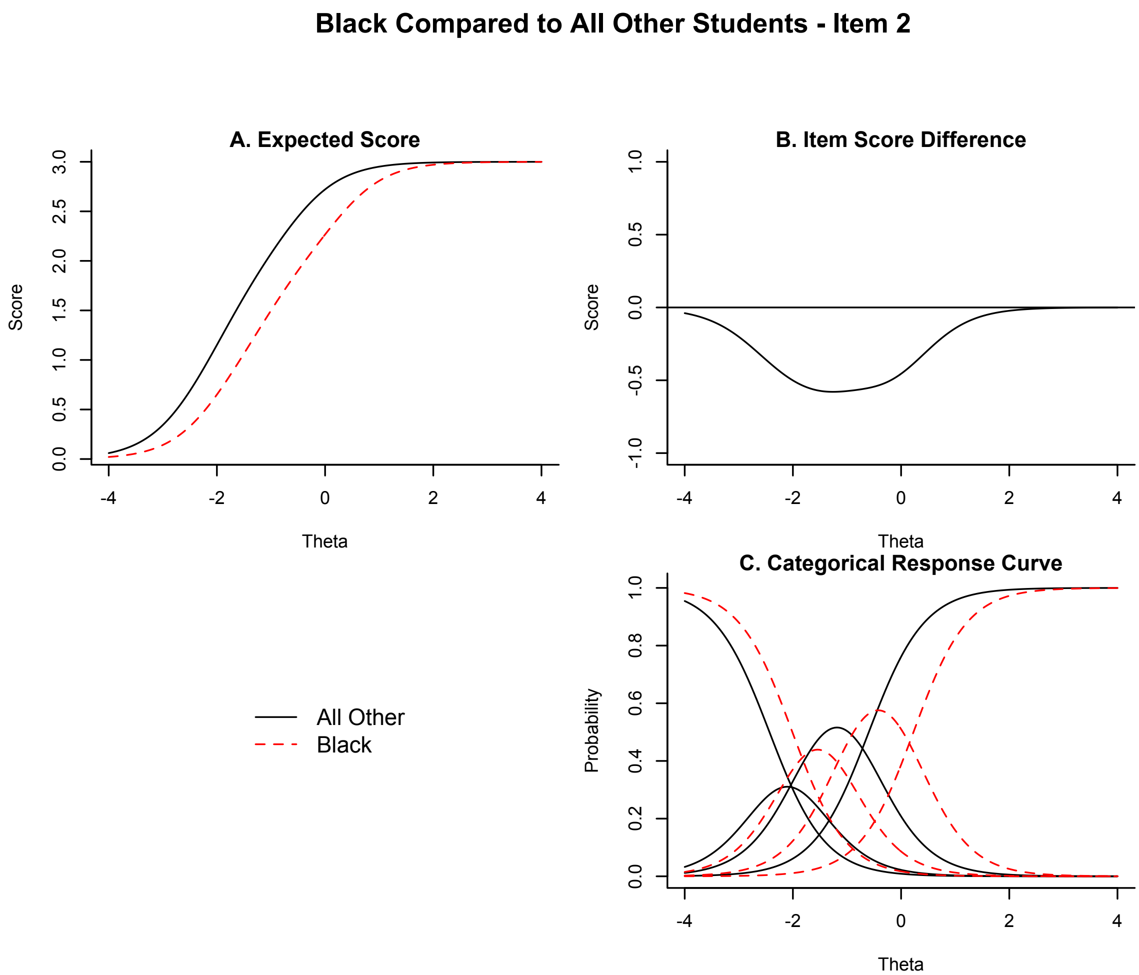


Figure D11

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for Black students on item 3.

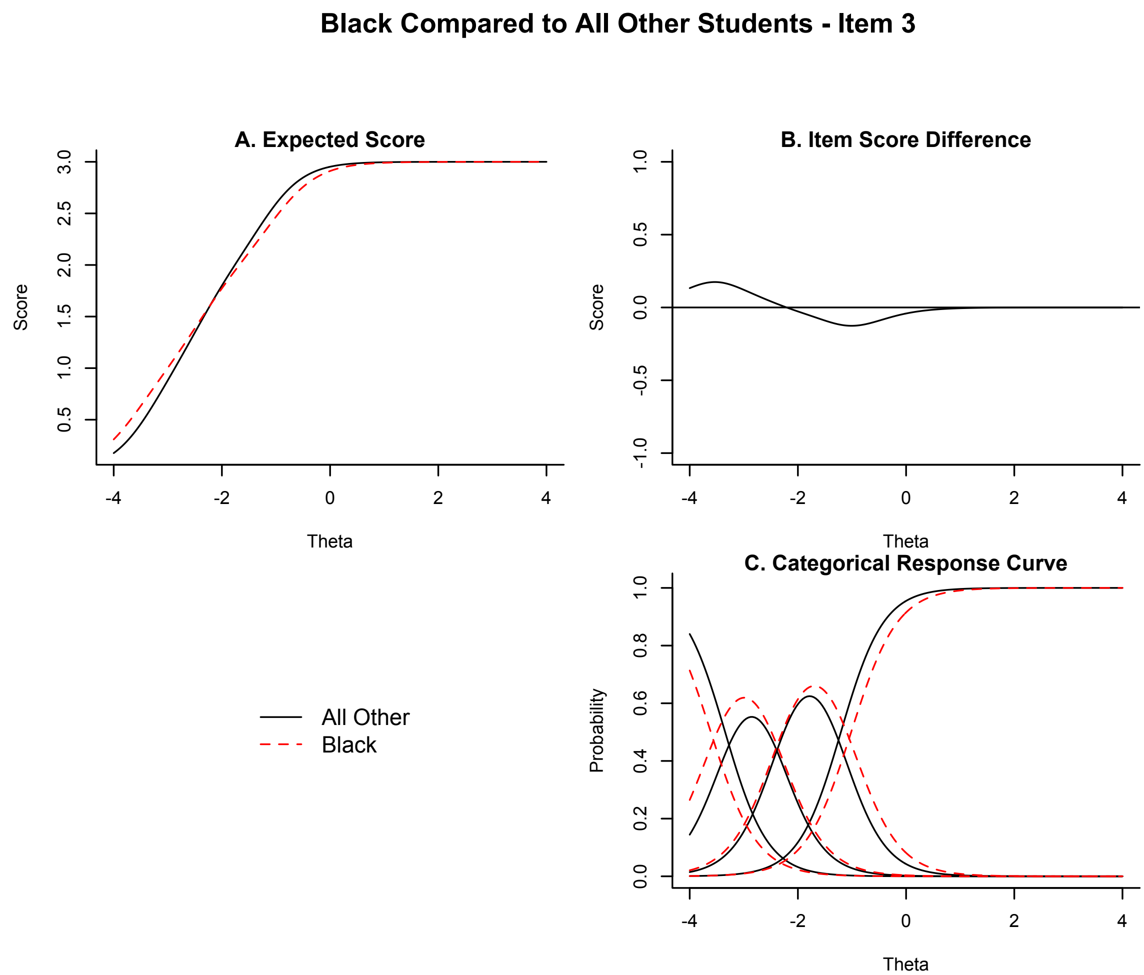


Figure D12

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for Black students on item 4.

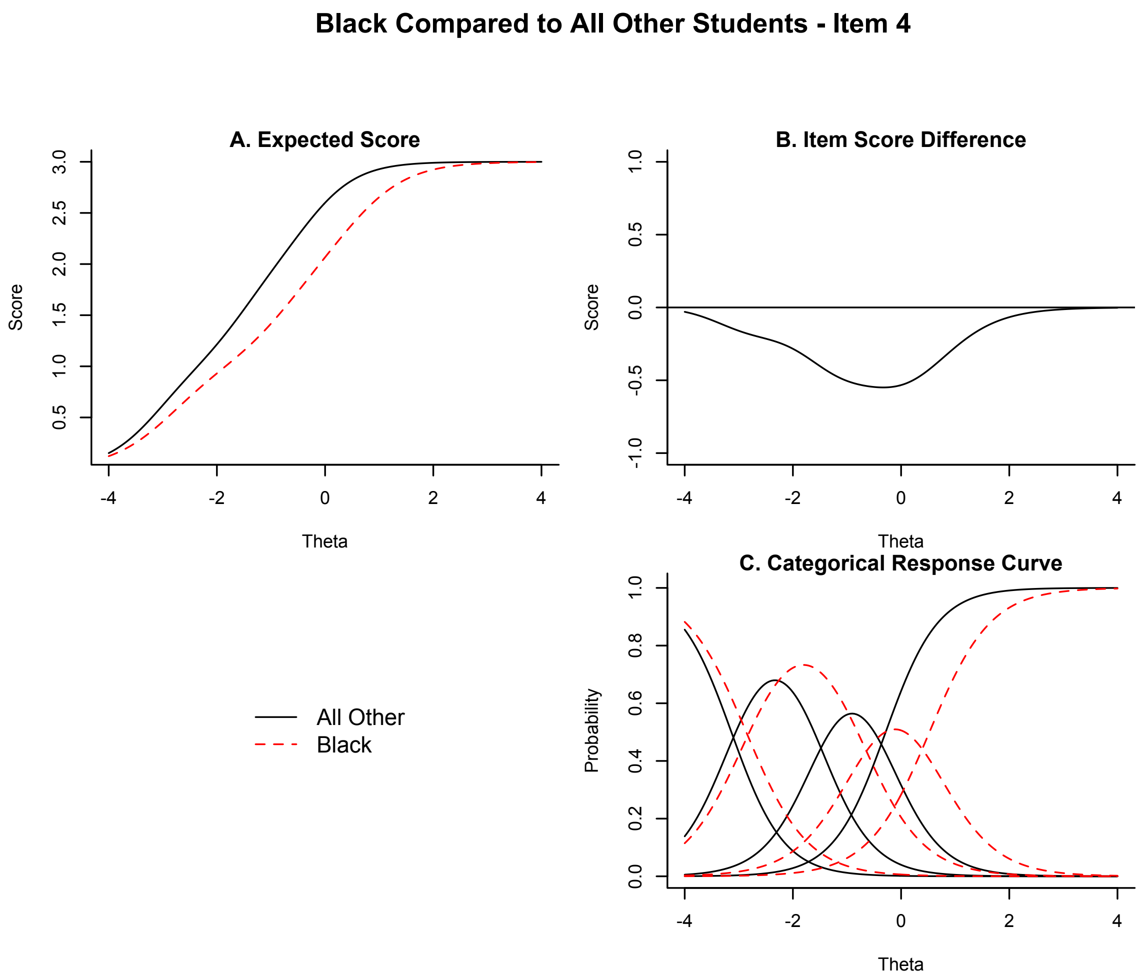


Figure D13

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for Black students on item 5.

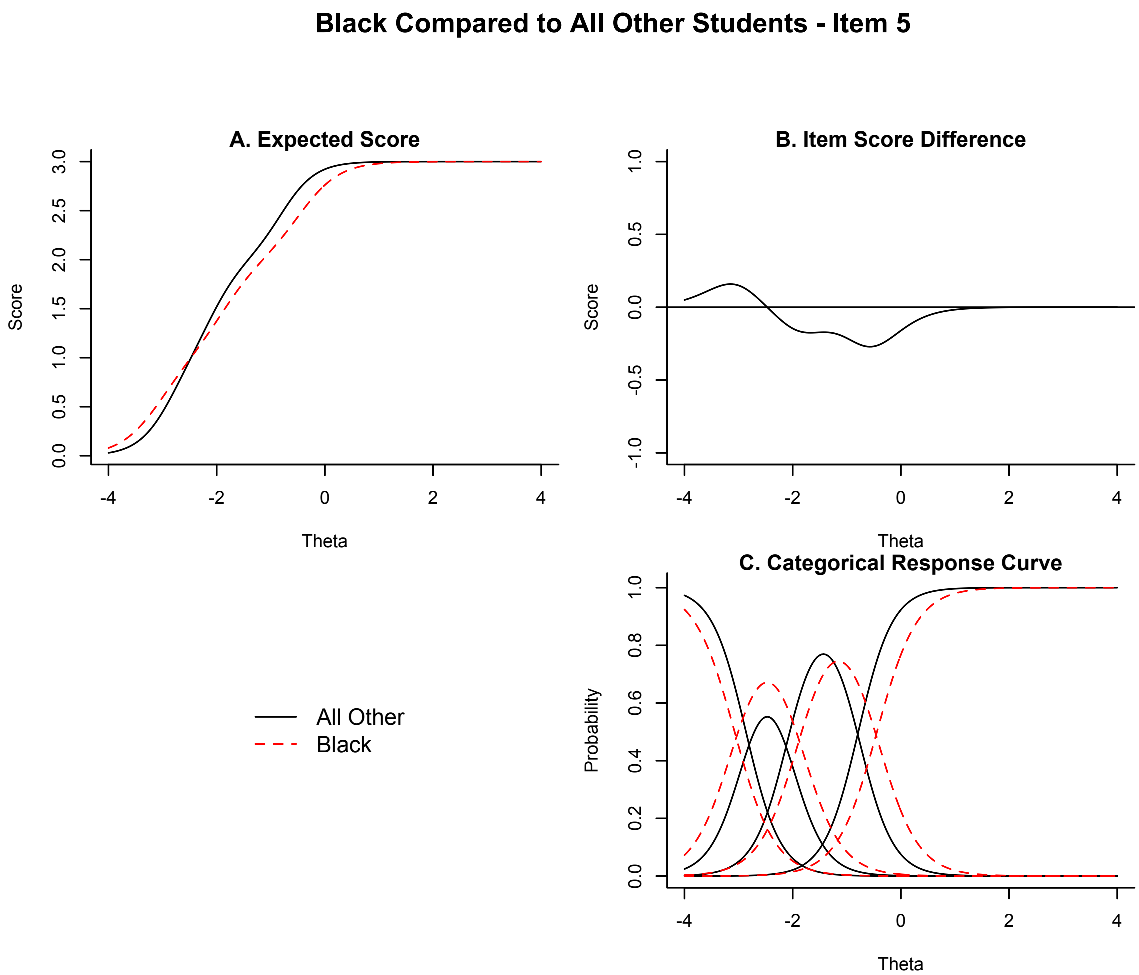


Figure D14

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for Black students on item 6.

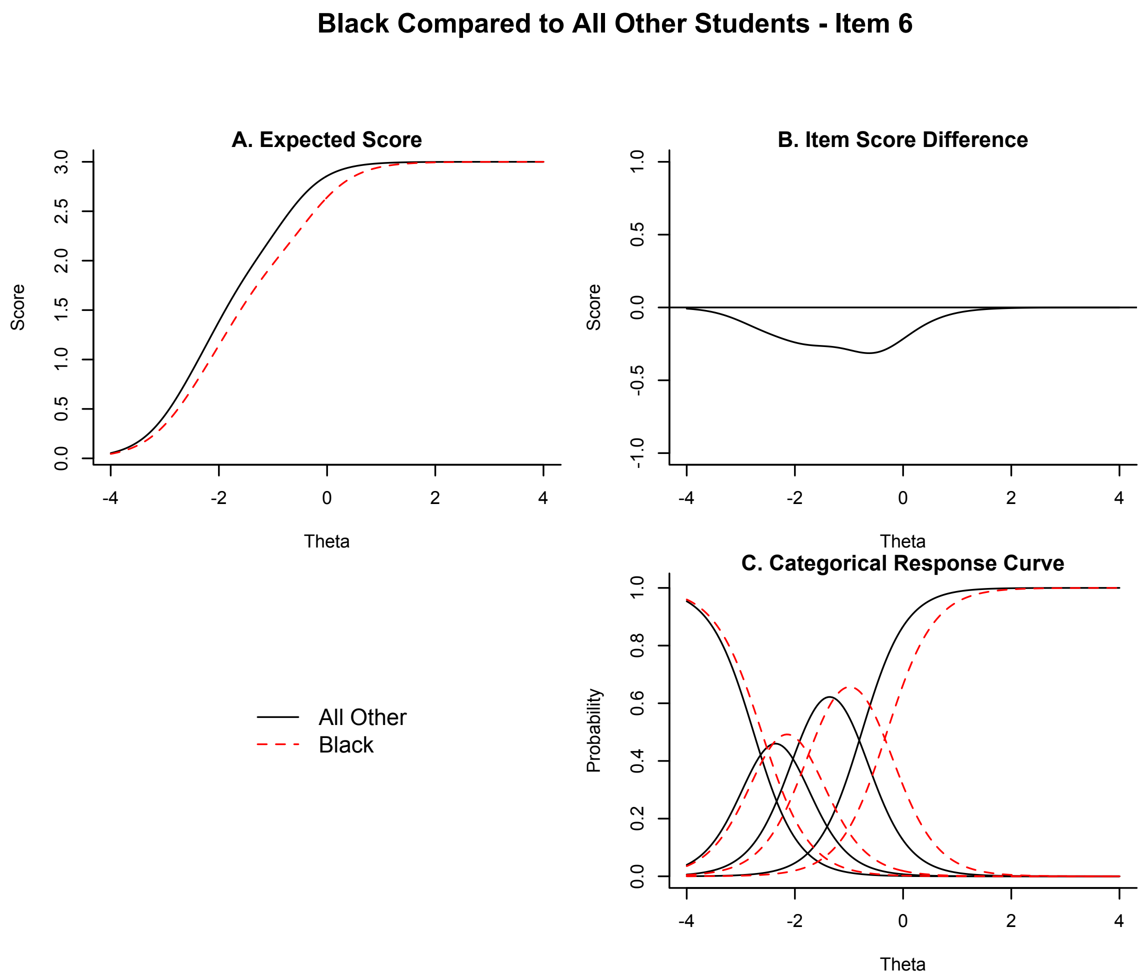


Figure D15

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for Black students on item 7.

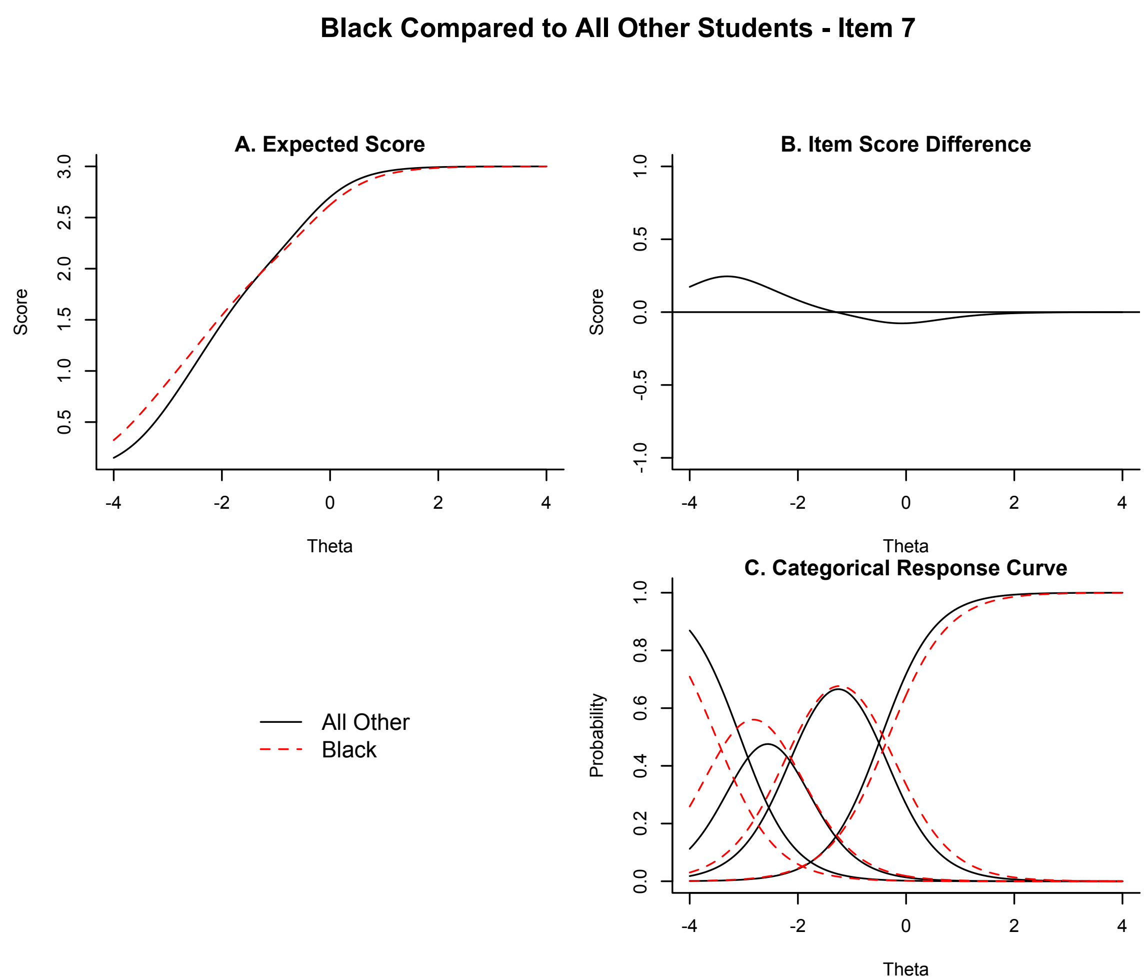


Figure D16

Graphs displaying the test response functions and the difference between the item response functions for Asian students.

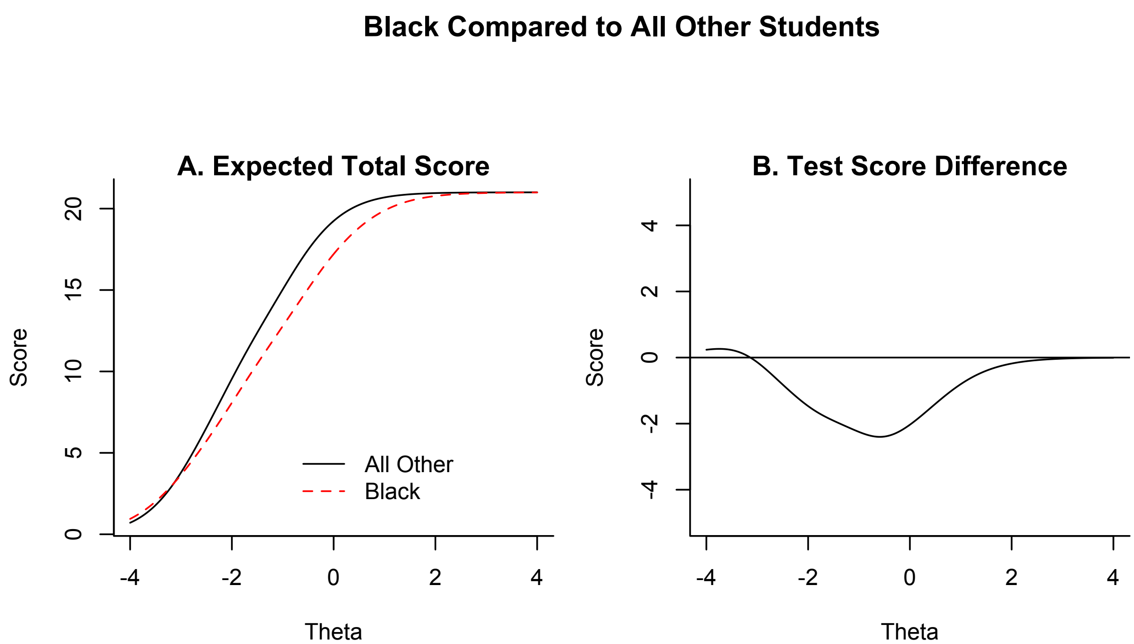


Figure D17

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for Hispanic students on item 1.

Hispanic Students

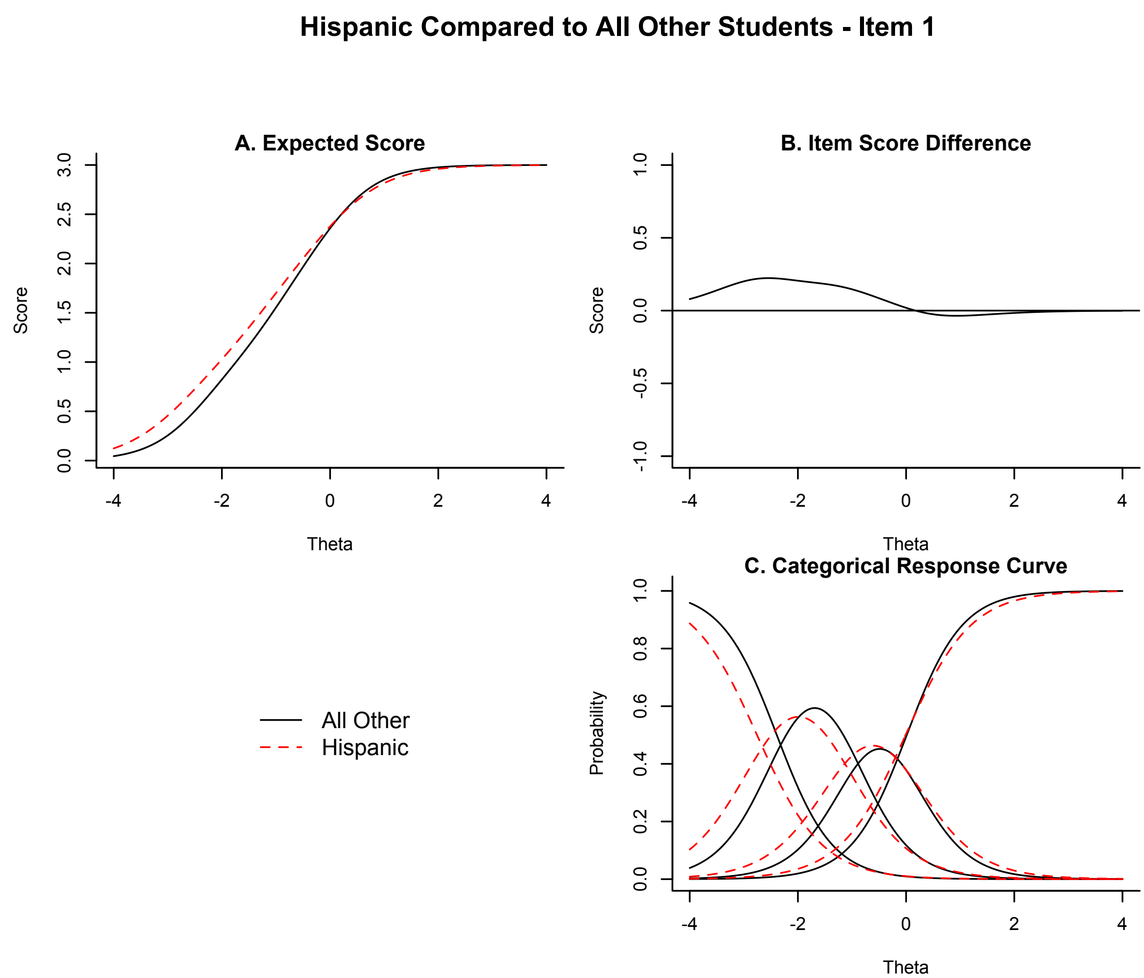


Figure D18

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for Hispanic students on item 2.

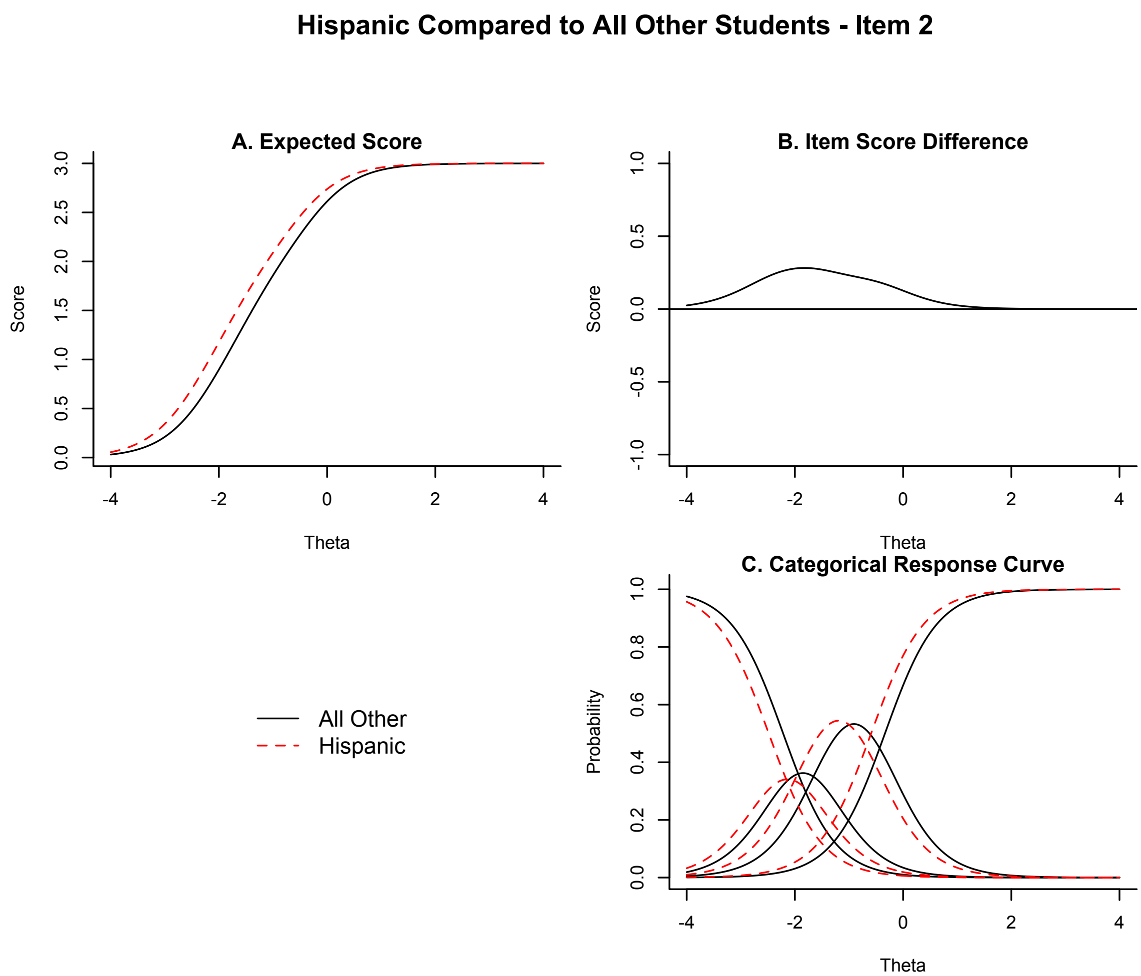


Figure D19

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for Hispanic students on item 3.

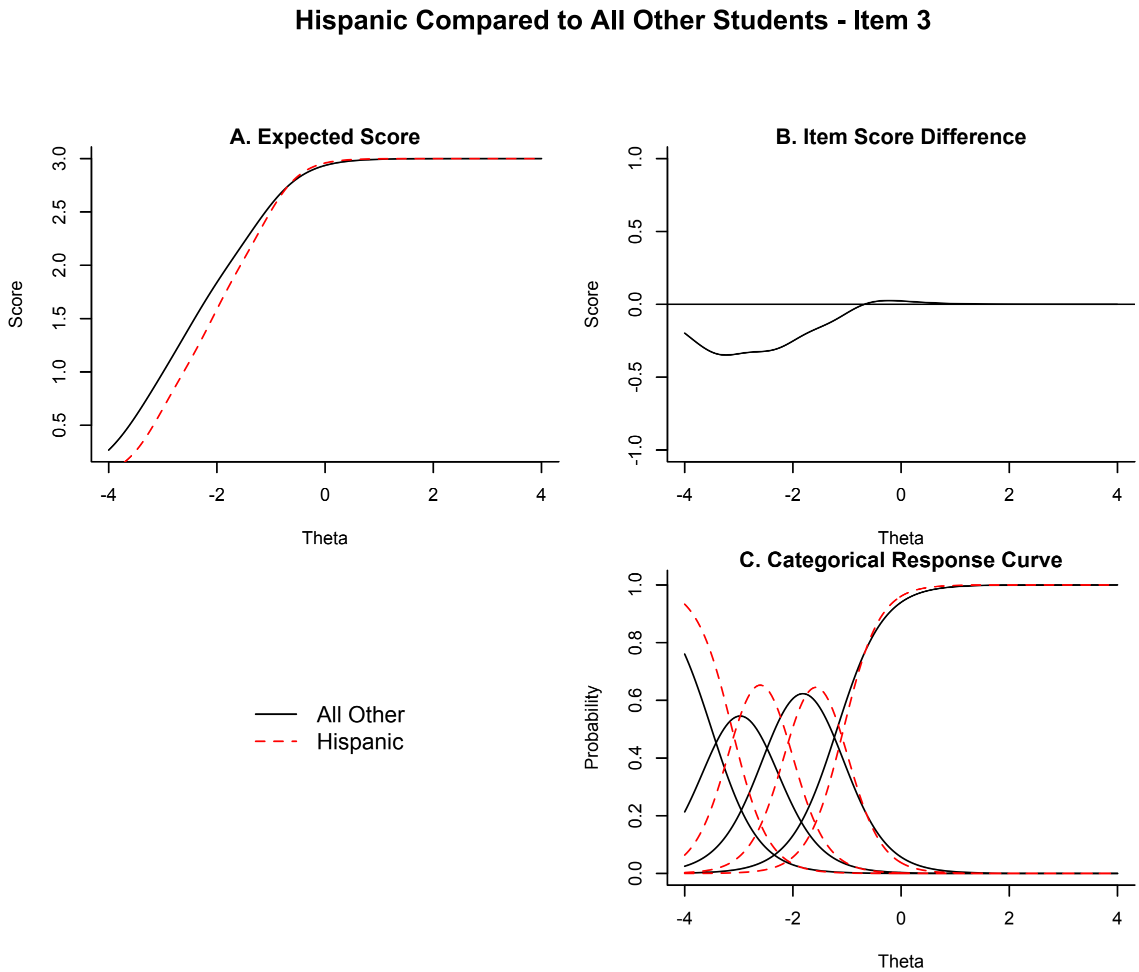


Figure D20

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for Hispanic students on item 4.

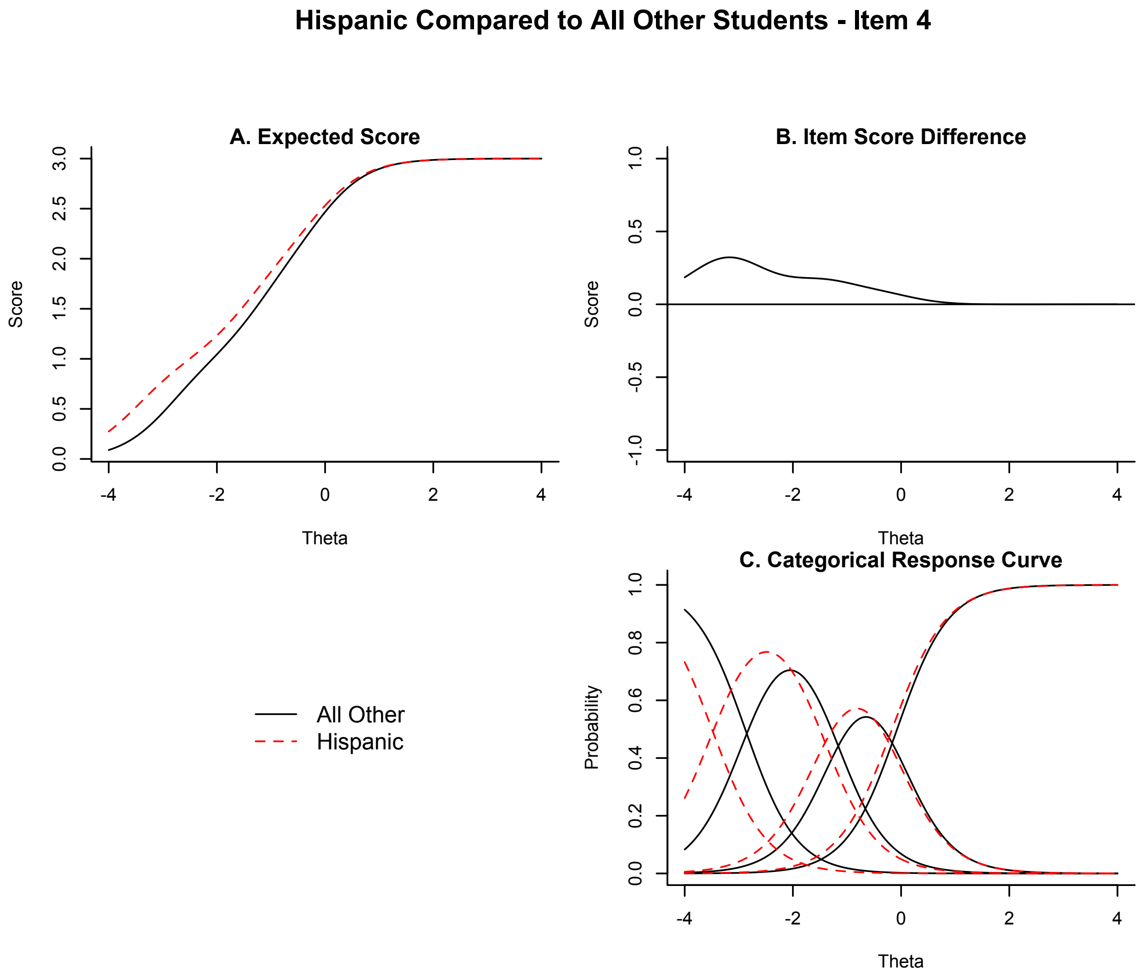


Figure D21

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for Hispanic students on item 5.

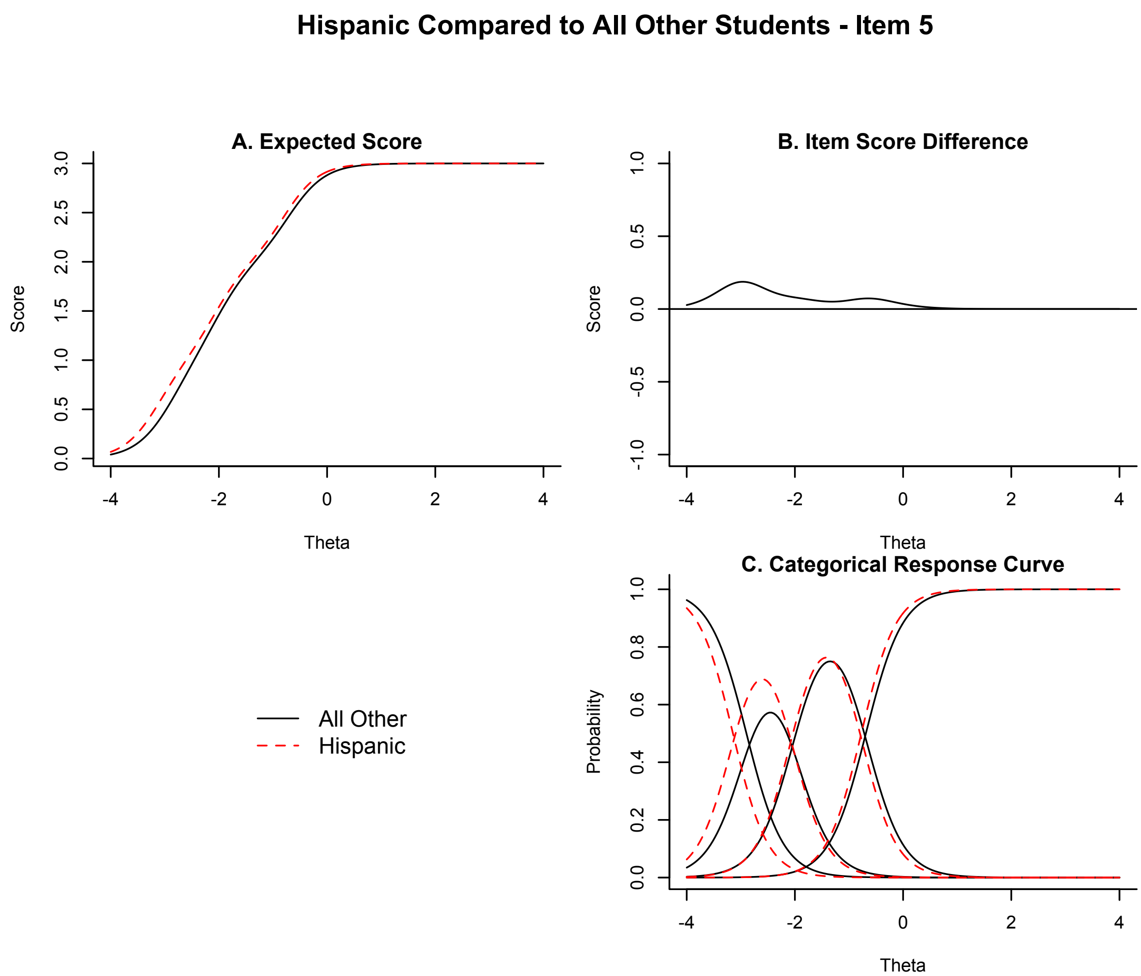


Figure D22

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for Hispanic students on item 6.

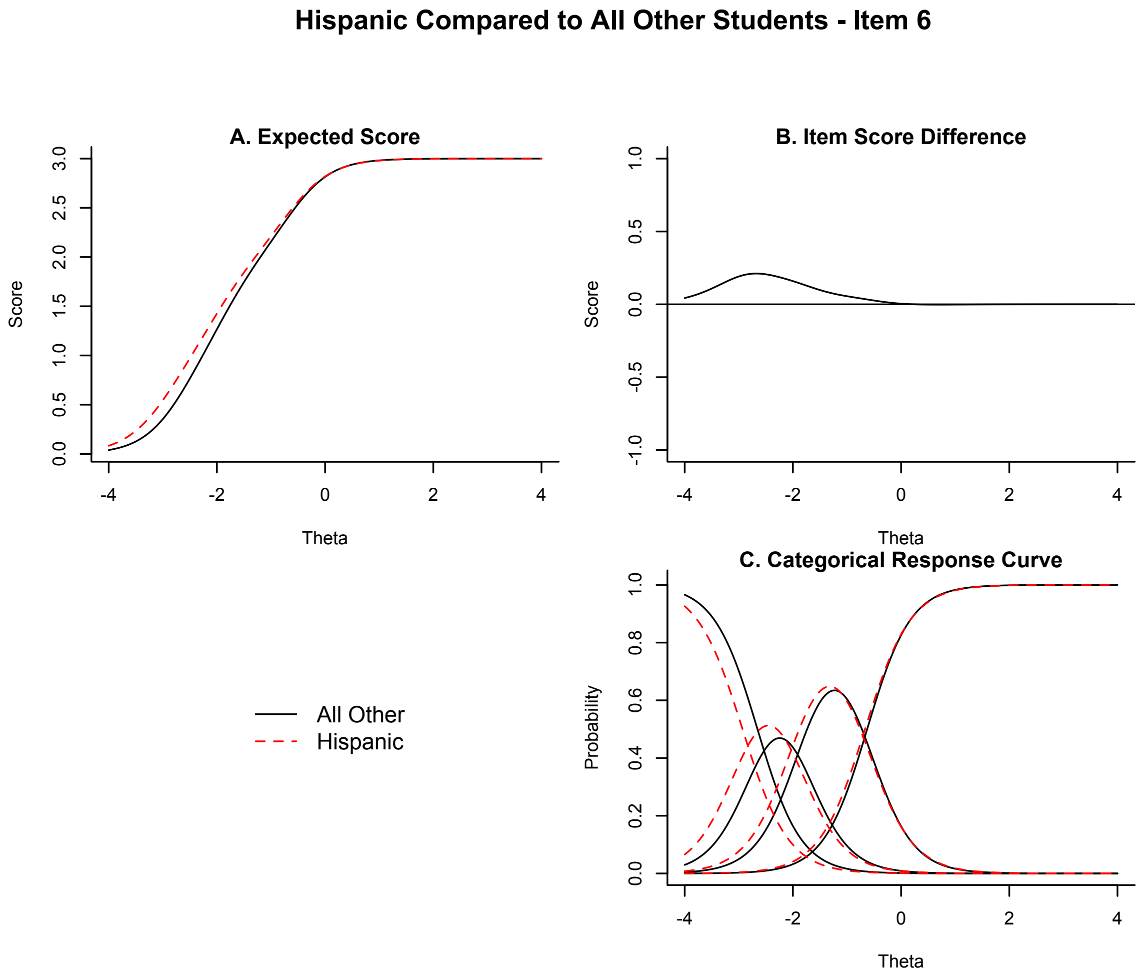


Figure D23

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for Hispanic students on item 7.

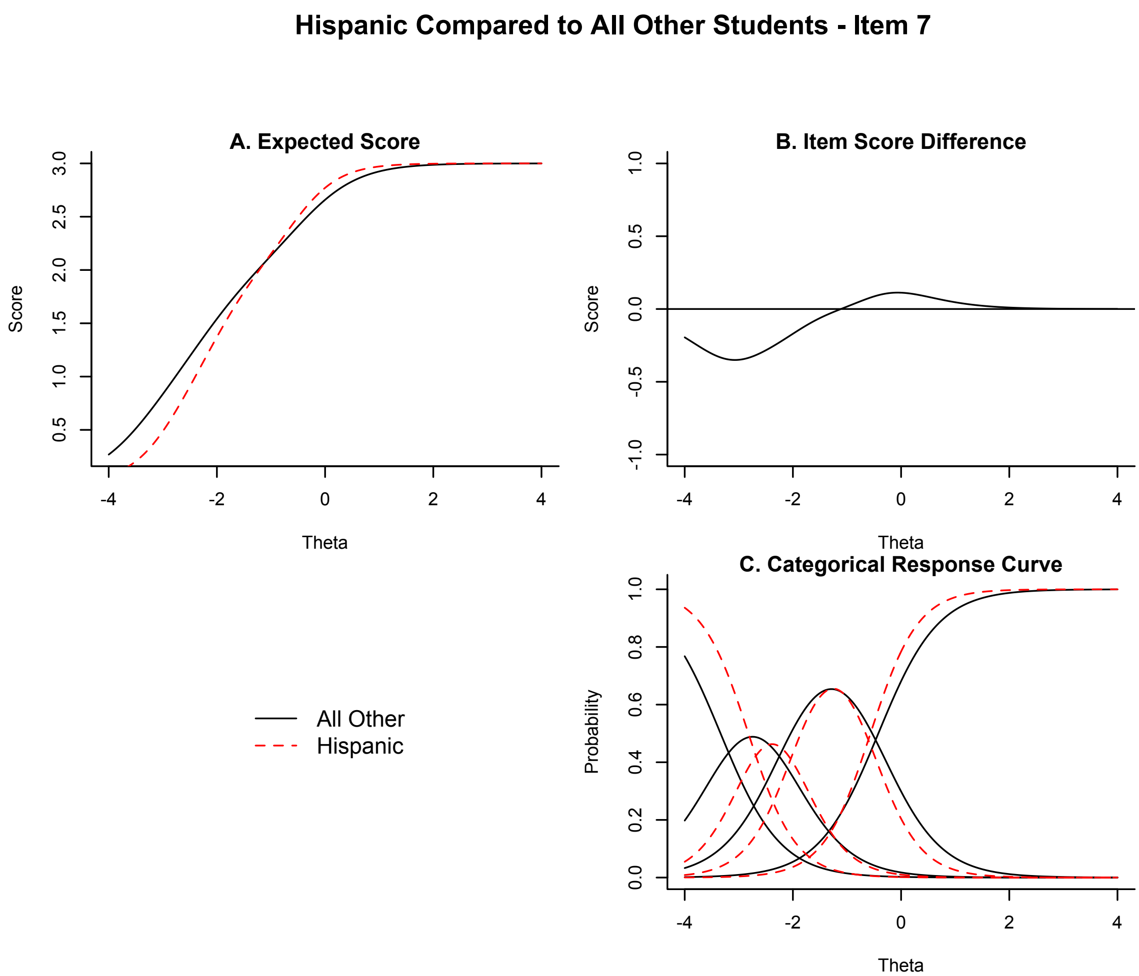


Figure D24

Graphs displaying the test response functions and the difference between the item response functions for Hispanic students.

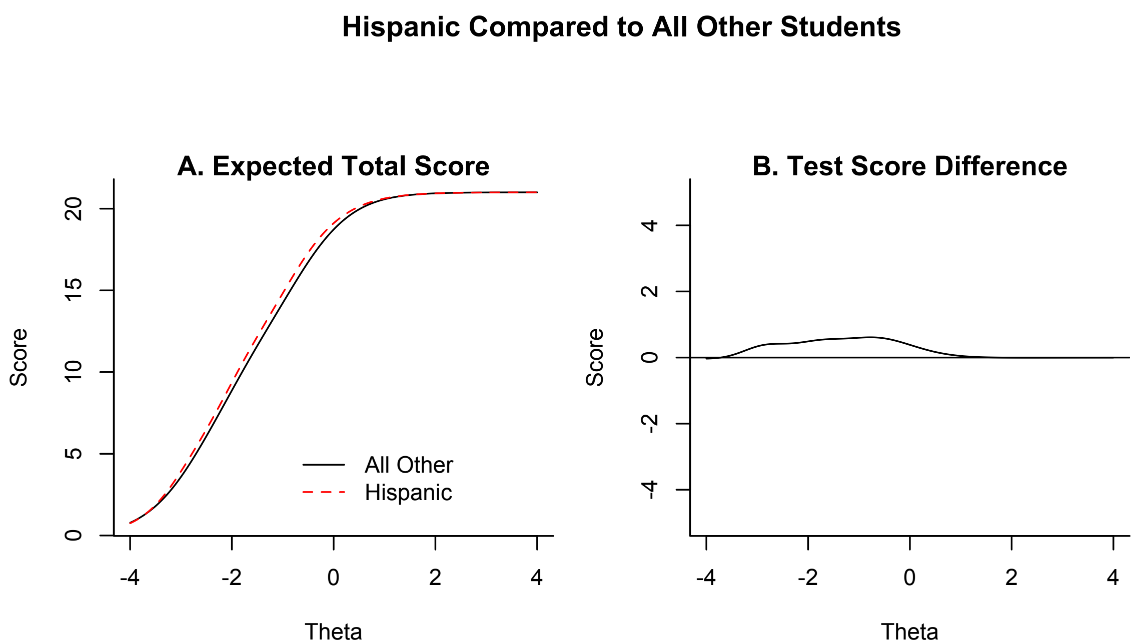


Figure D25

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for students with multiple races and ethnicities on item 1.

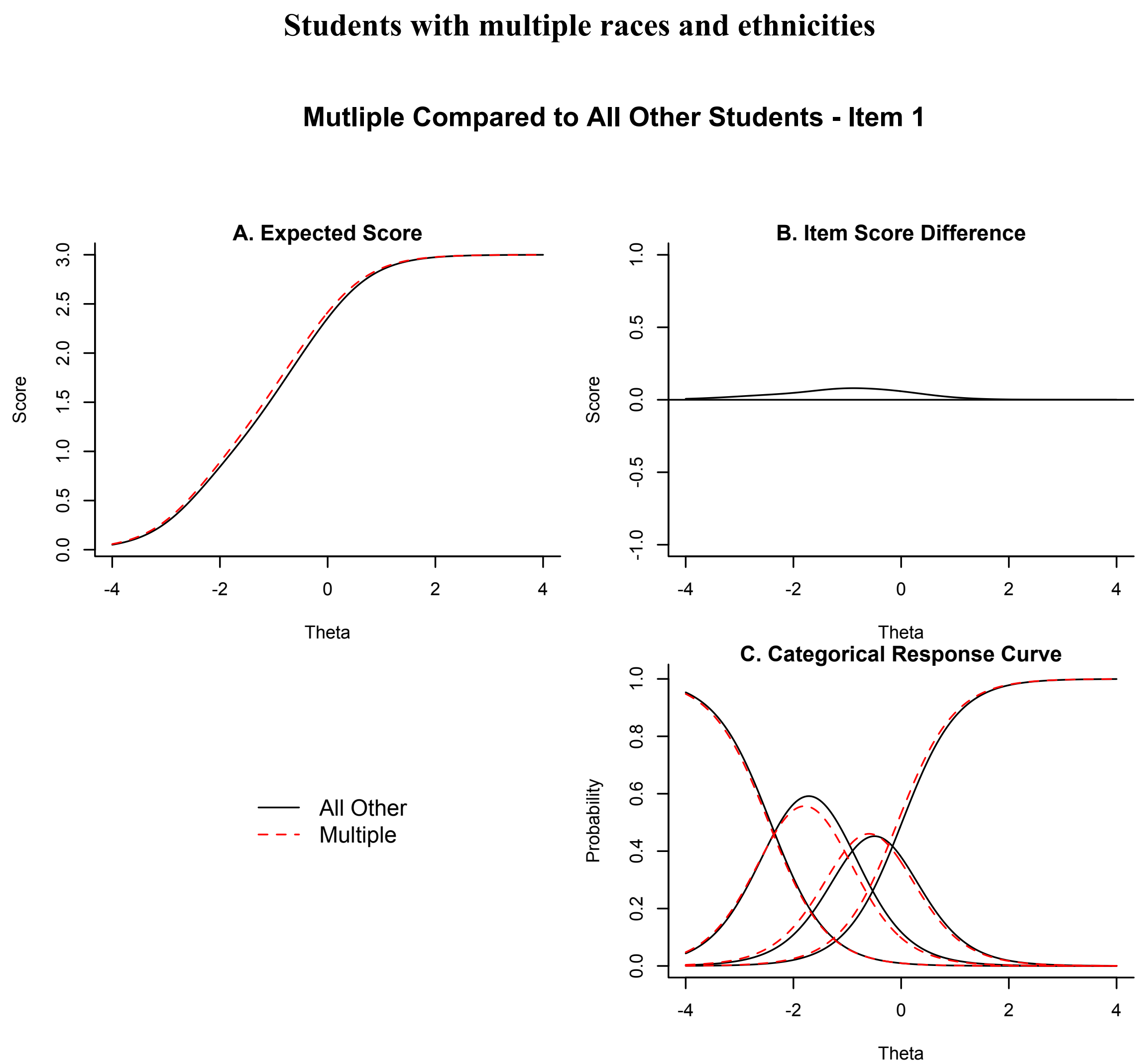


Figure D26

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for students with multiple races and ethnicities on item 2.

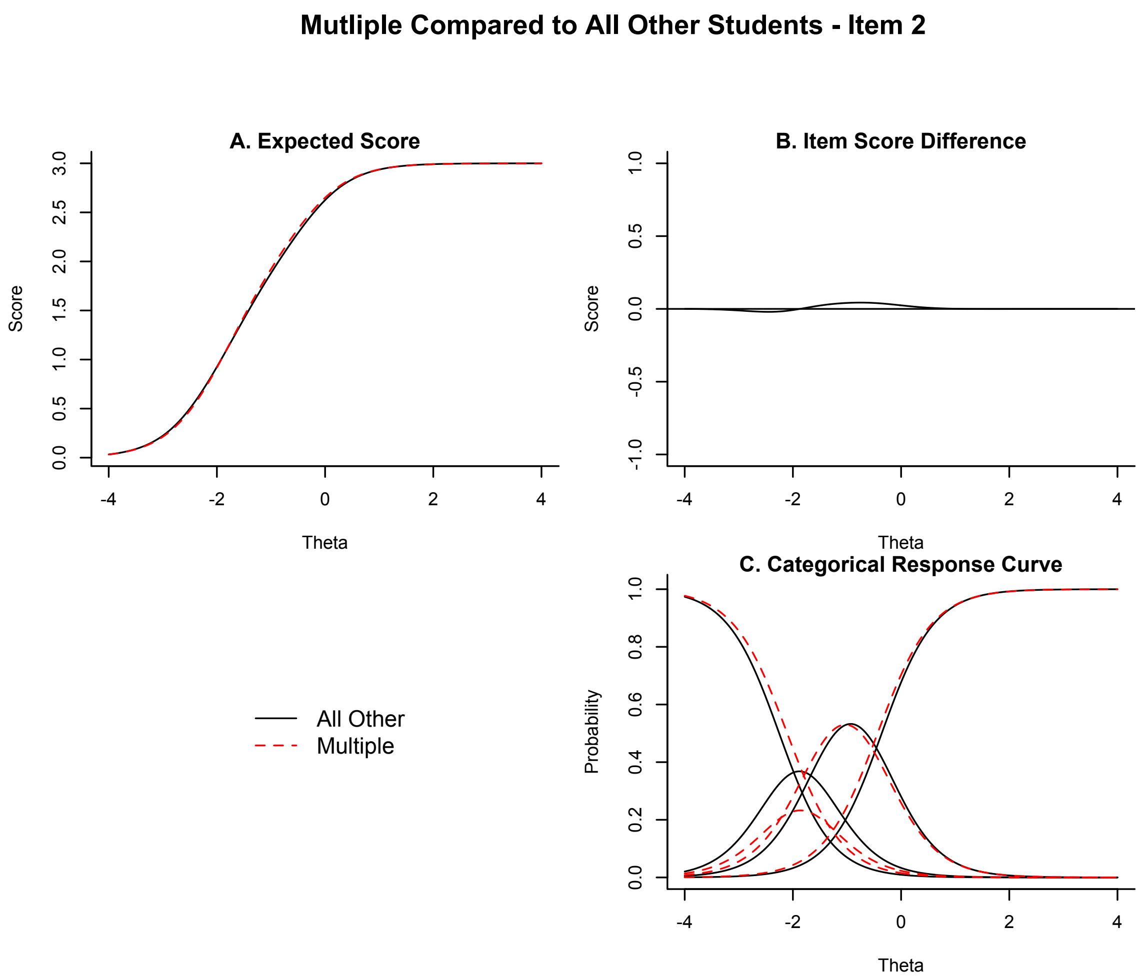


Figure D27

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for students with multiple races and ethnicities on item 3.

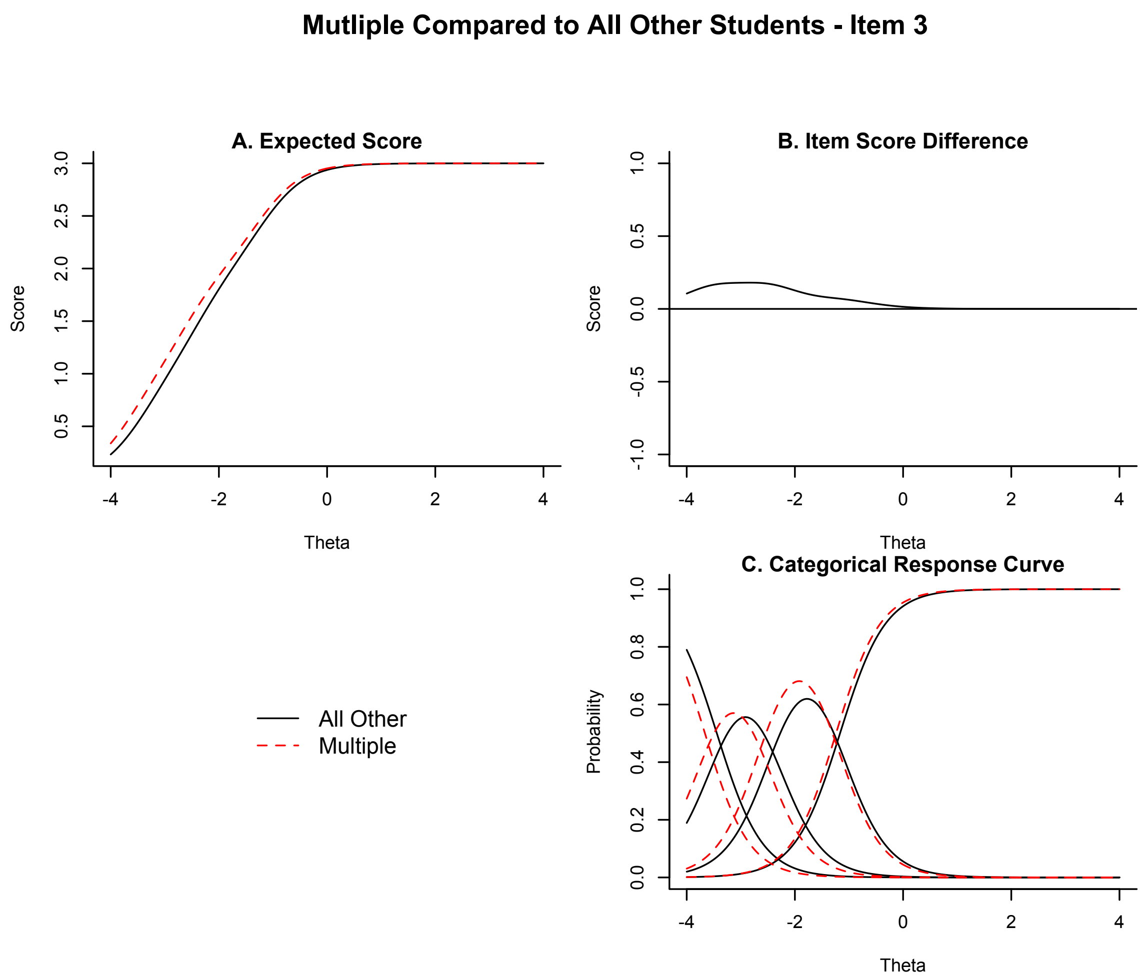


Figure D28

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for students with multiple races and ethnicities on item 4.

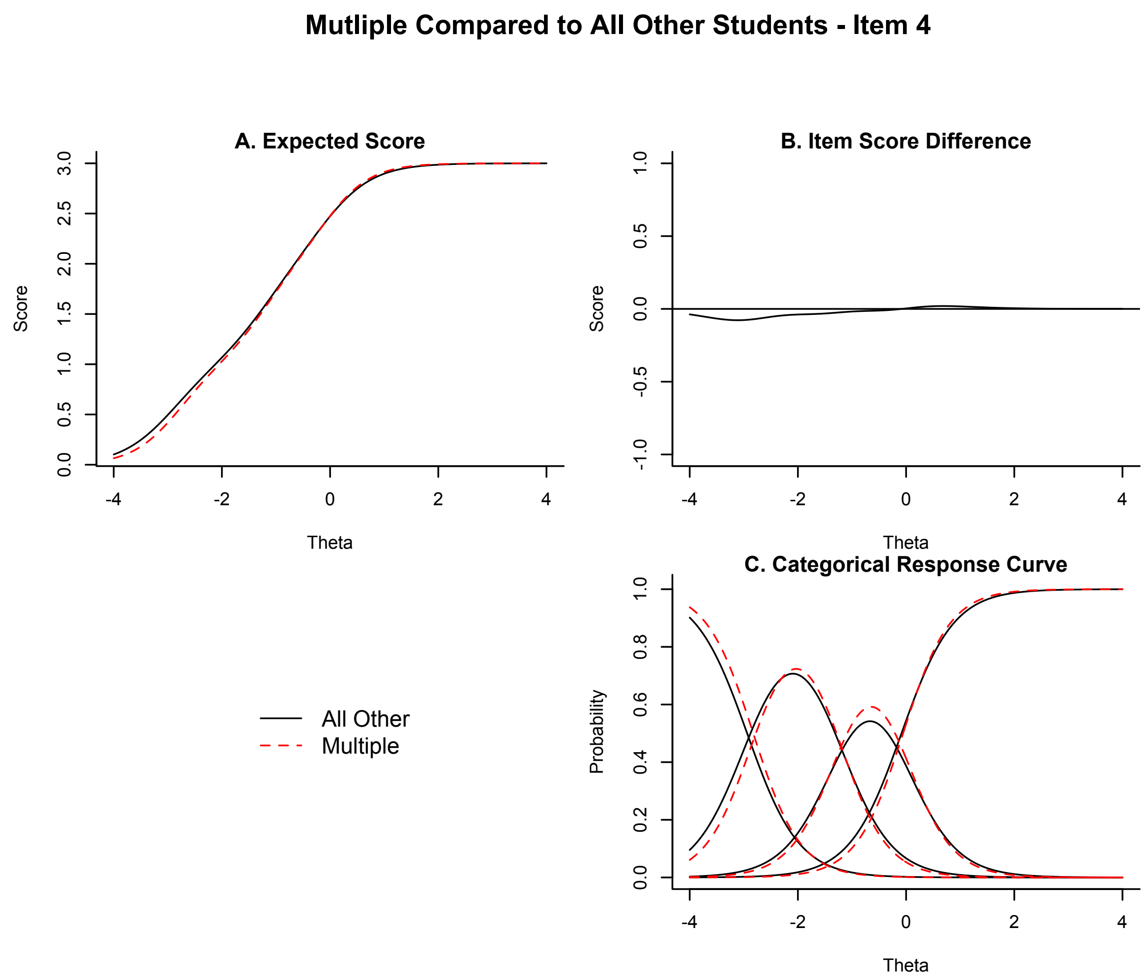


Figure D29

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for students with multiple races and ethnicities on item 5.

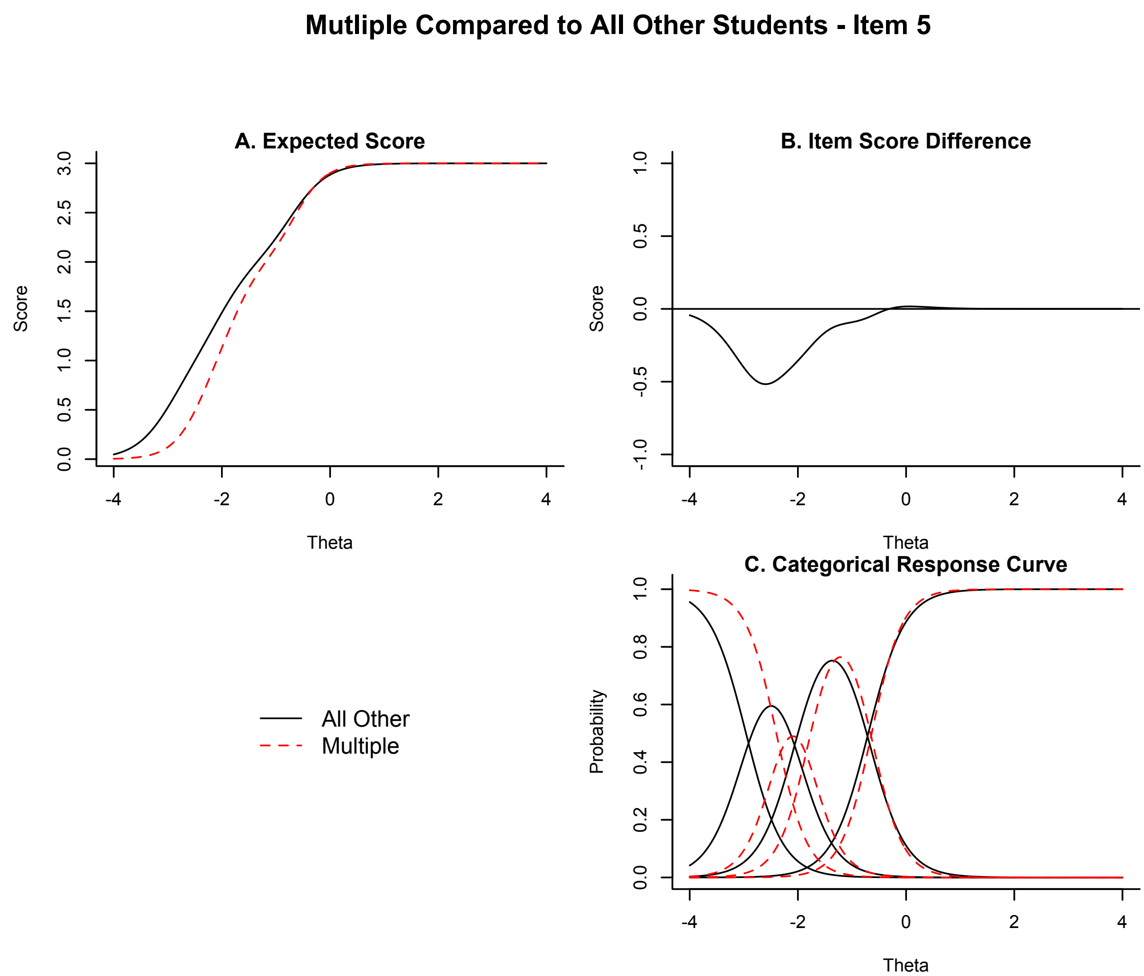


Figure D30

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for students with multiple races and ethnicities on item 6.

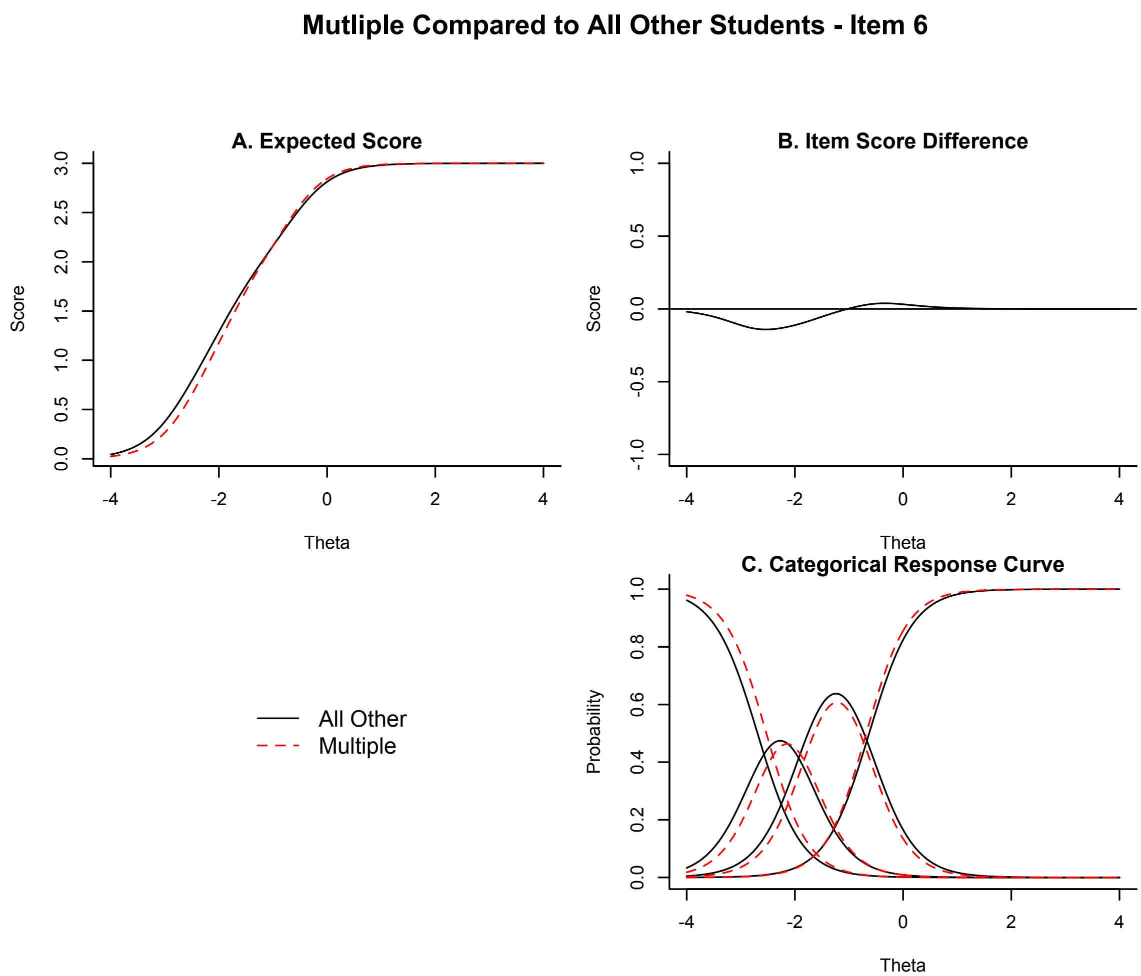


Figure D31

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for students with multiple races and ethnicities on item 7.

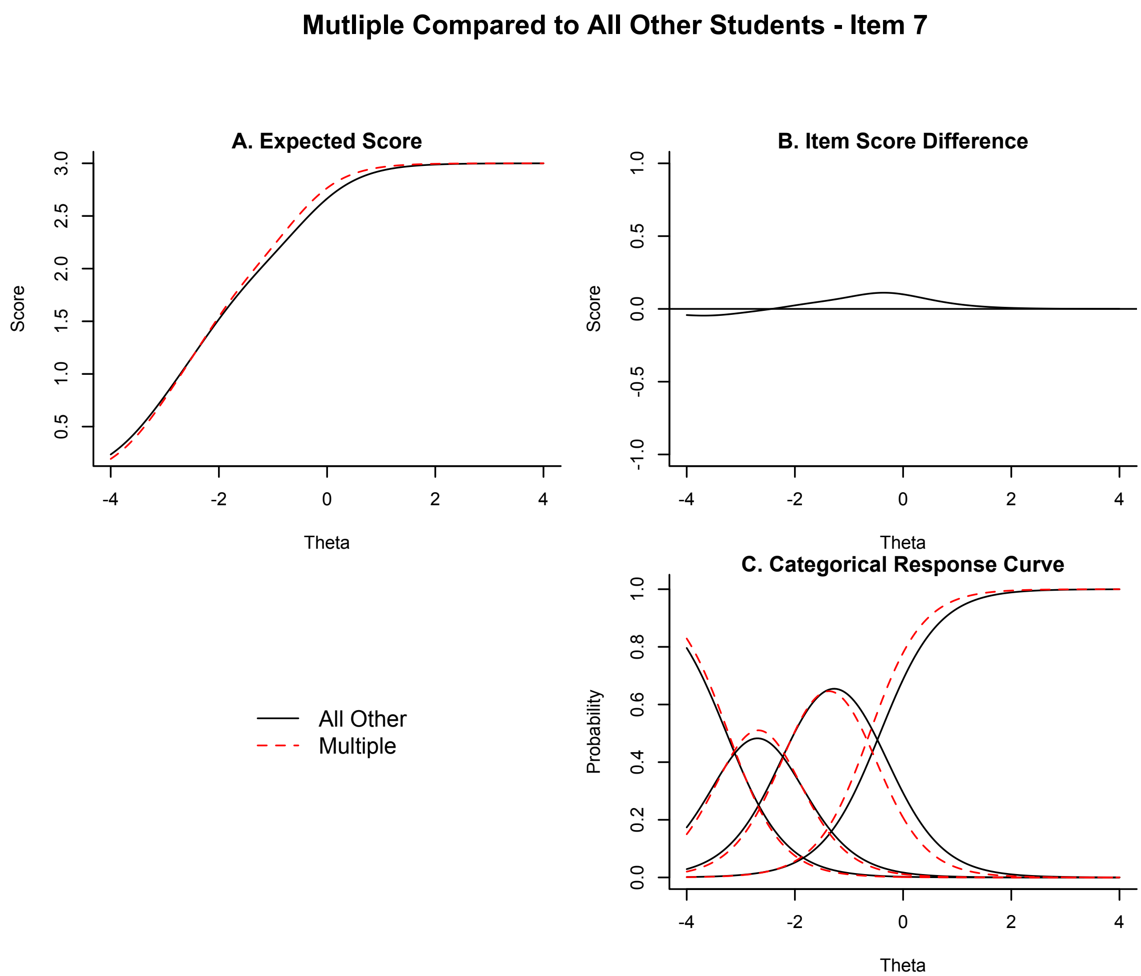


Figure D32

Graphs displaying the test response functions and the difference between the item response functions for students with multiple races and ethnicities.

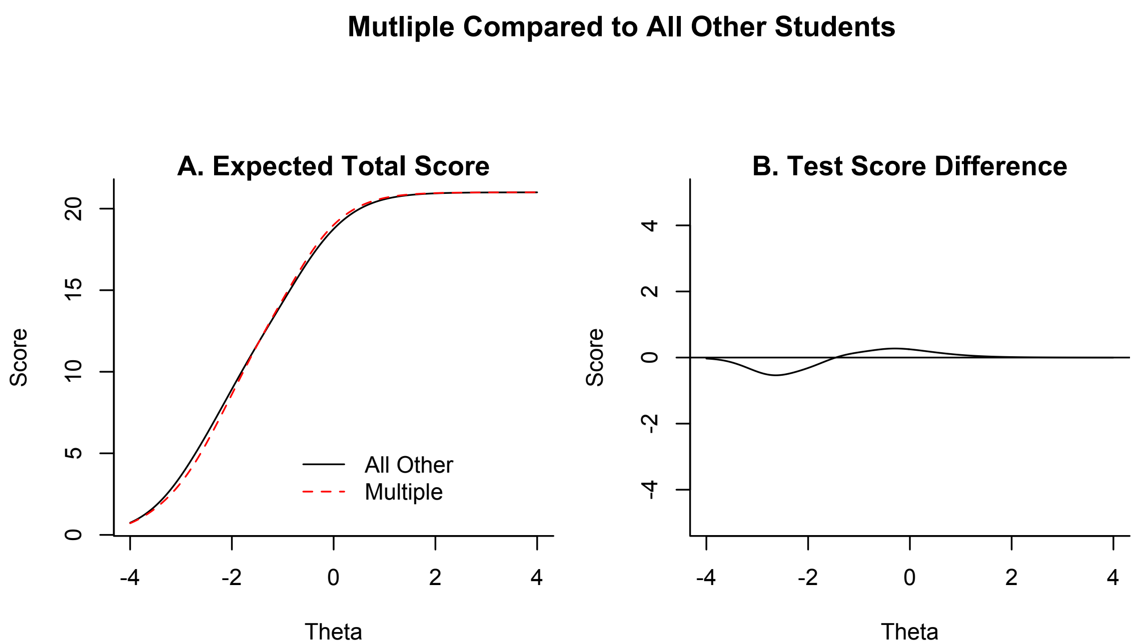


Figure D33

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for White students on item 1.

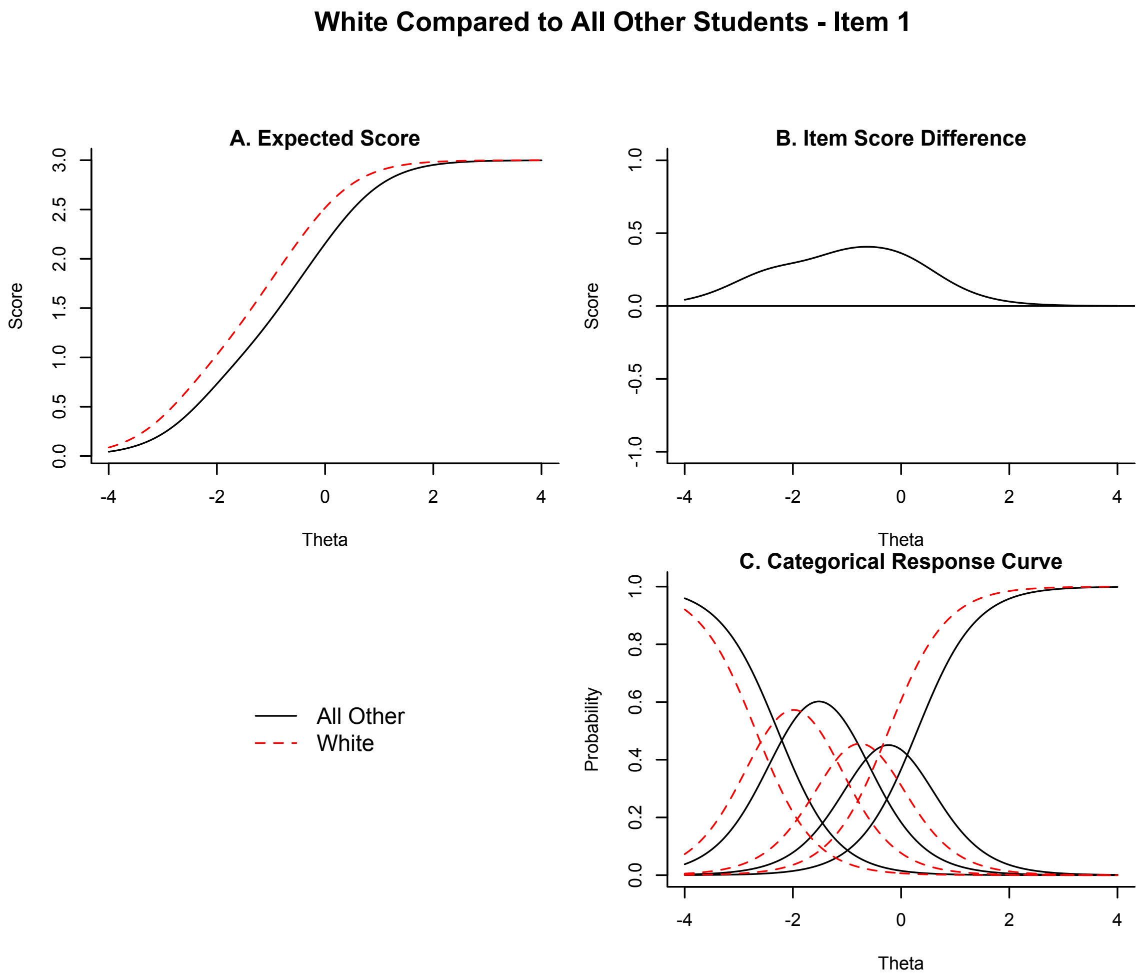


Figure D34

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for White students on item 2.

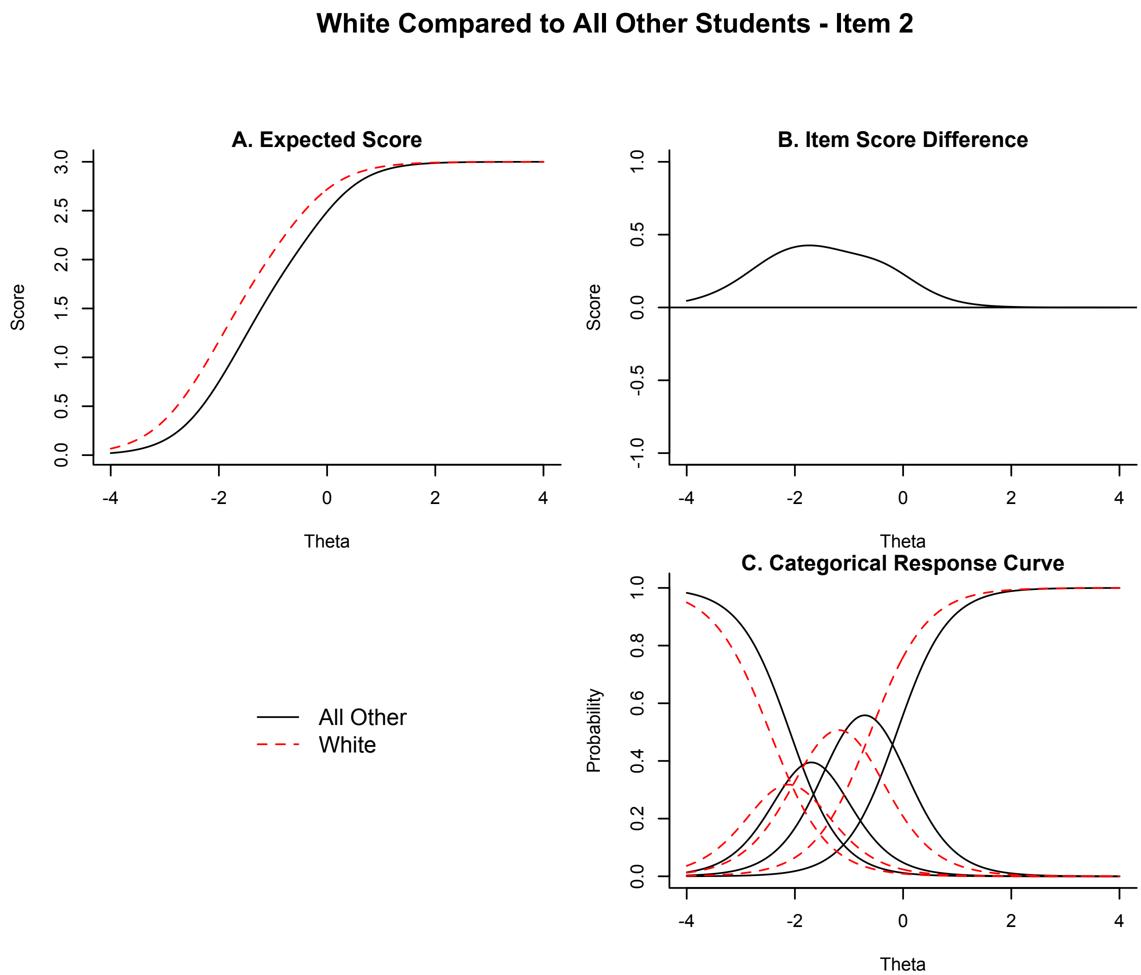


Figure D35

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for White students on item 3.

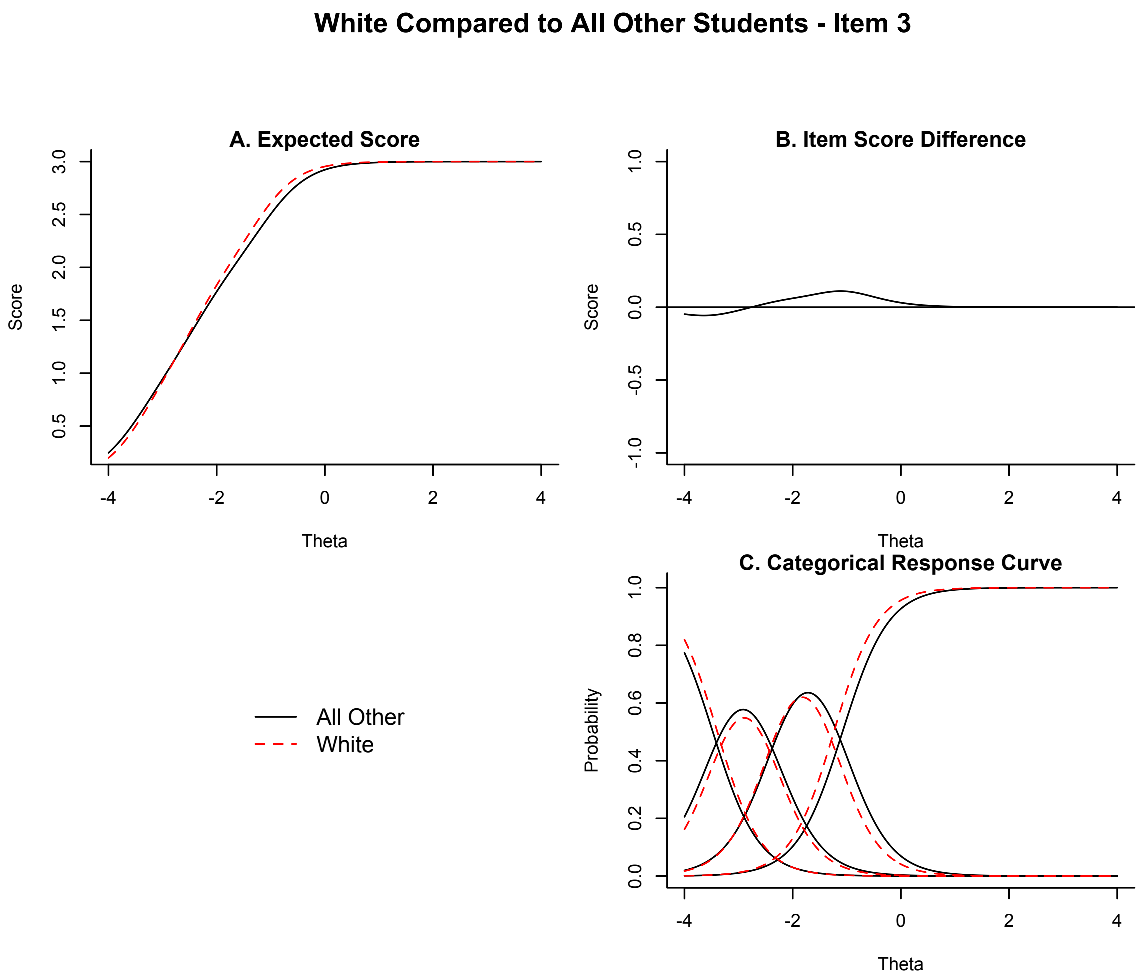


Figure D36

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for White students on item 4.

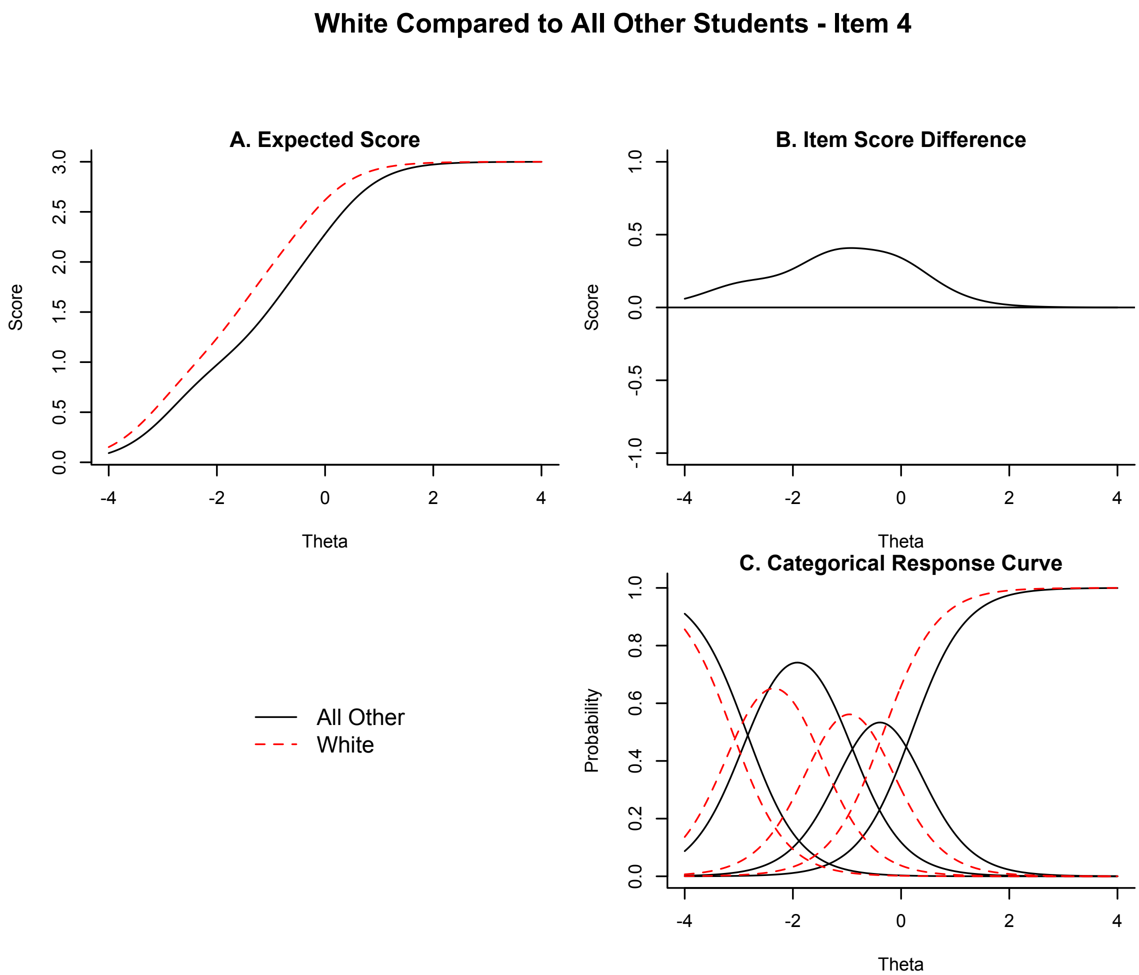


Figure D37

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for White students on item 5.

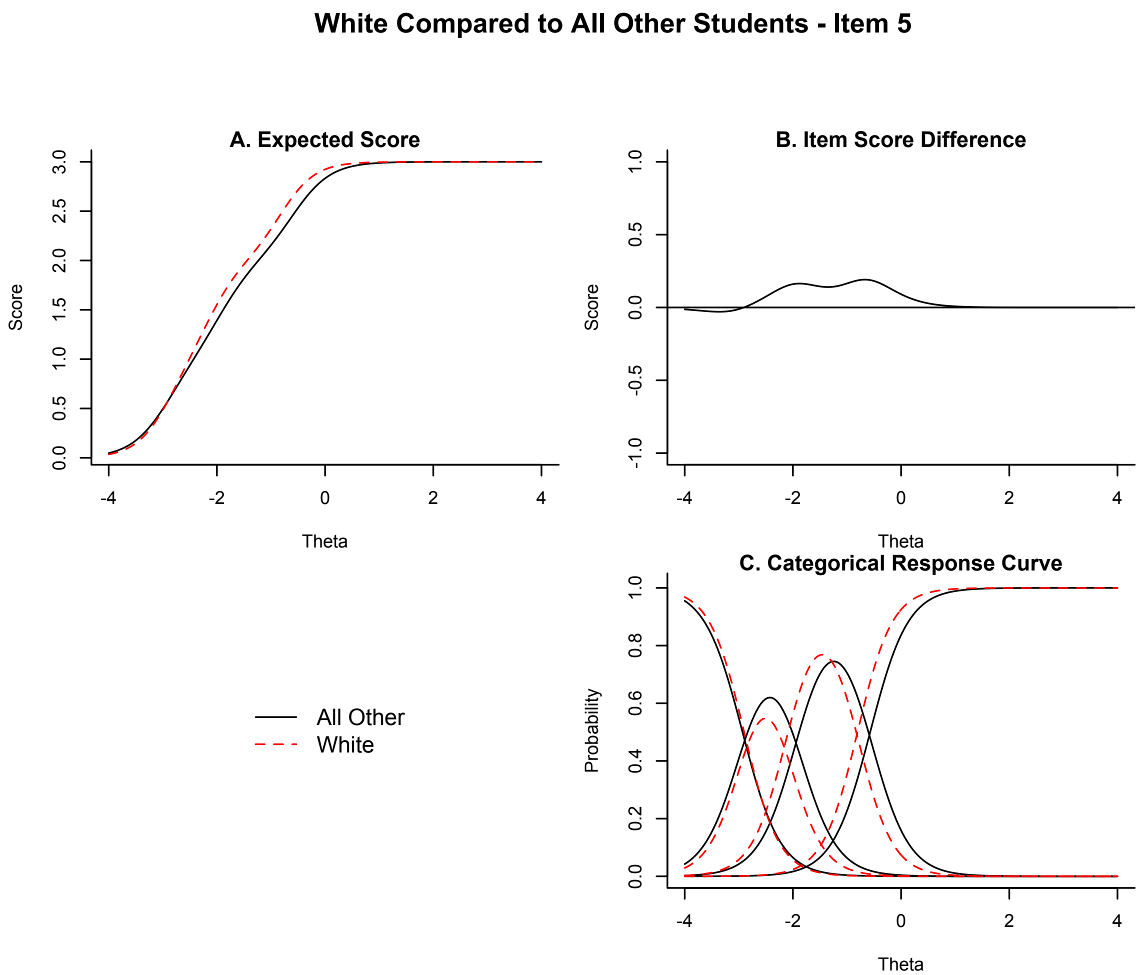


Figure D38

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for White students on item 6.

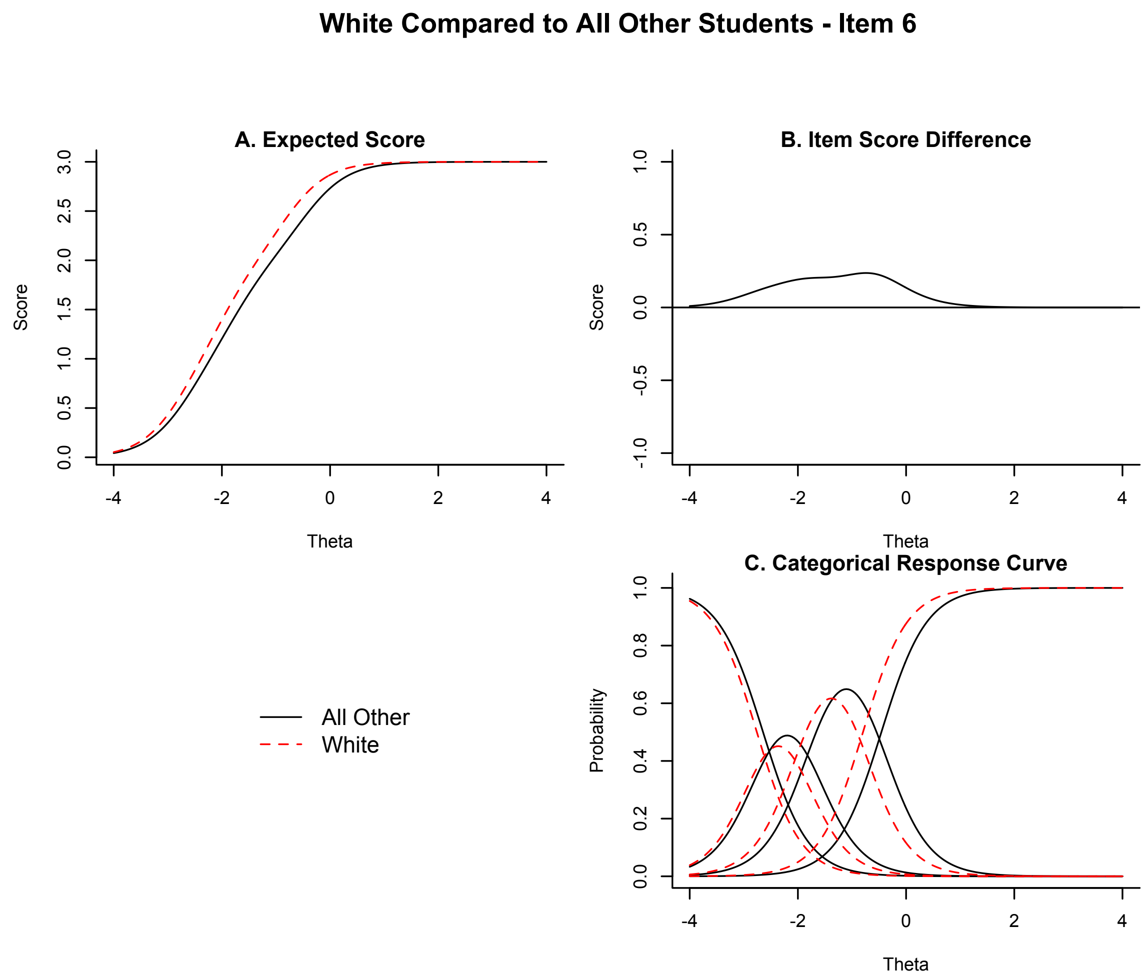


Figure D39

Graphs displaying the item response functions, the difference between the item response functions, and the categorical response curve for White students on item 7.

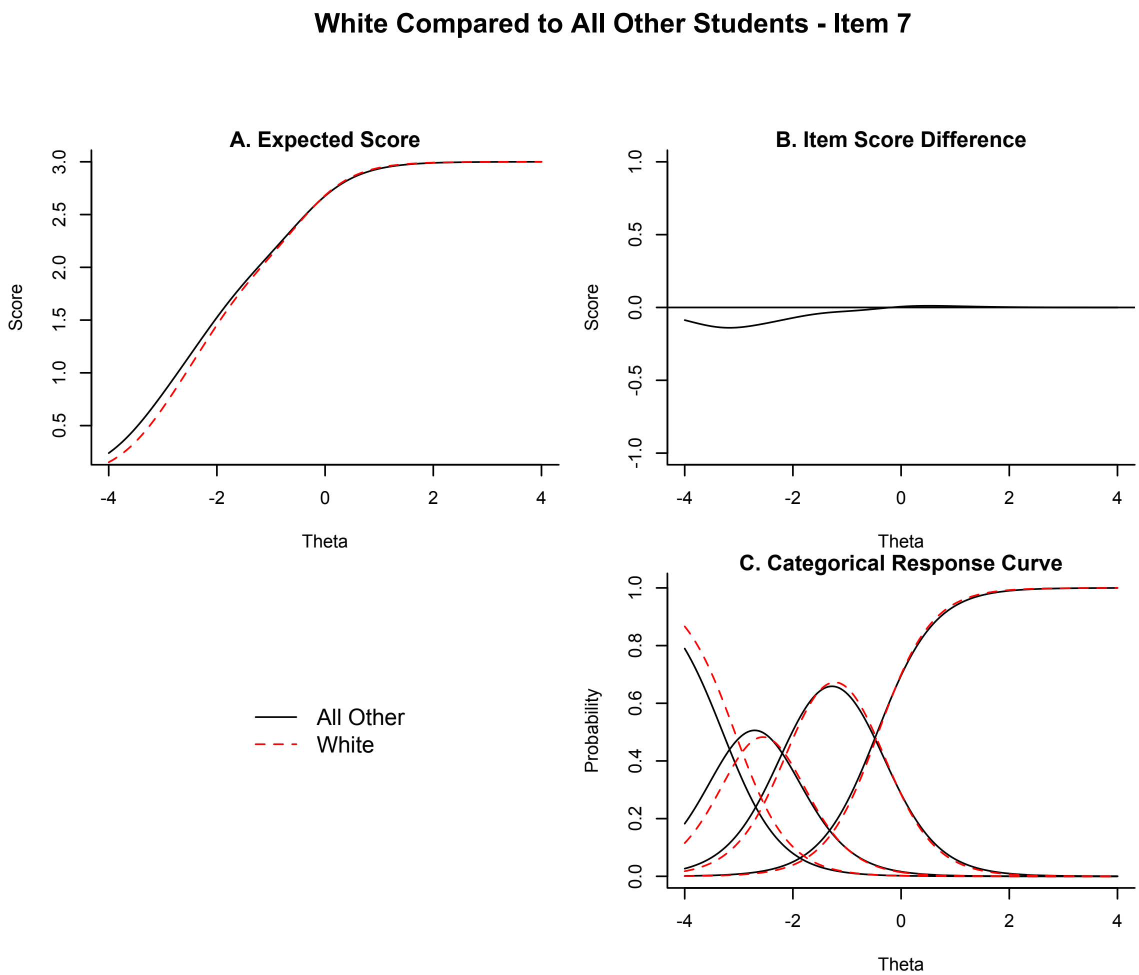
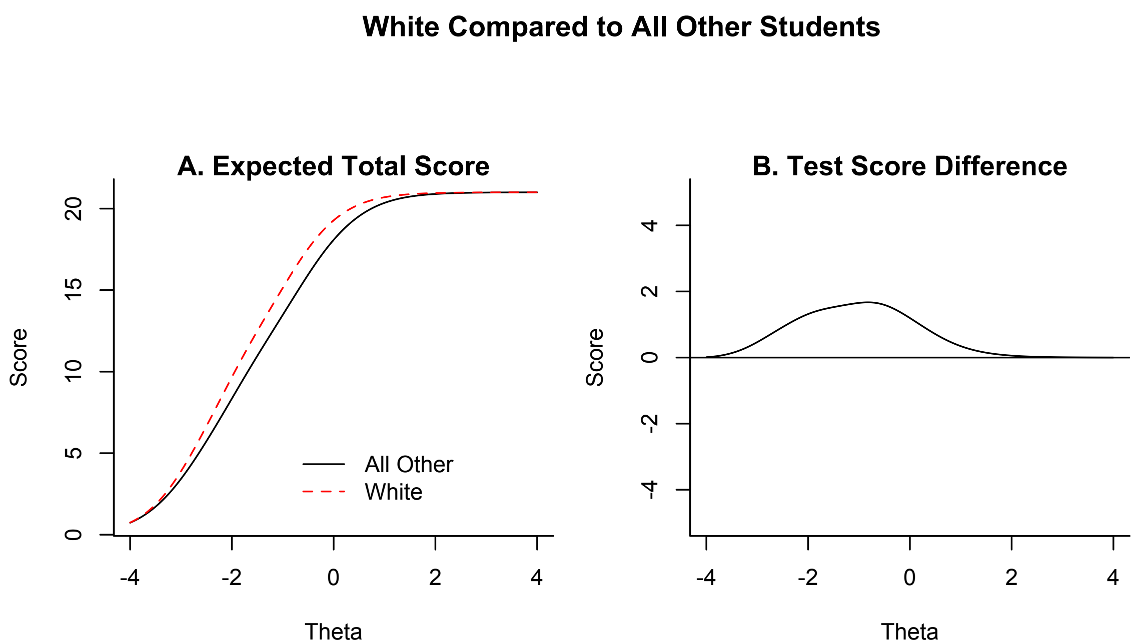


Figure D40

Graphs displaying the test response functions and the difference between the item response functions for White students.



VITA

Jared Izumi was born in Los Angeles, CA on July 11, 1988 to Michael and Debra Izumi. After graduating from St. Francis High School in 2006, Jared studied at the University of California, Berkeley. Jared graduated with his Bachelor of Arts degree in psychology and minor in education in May 2011. He worked as a reading specialist and college admissions consultant for two years before returning to graduate school. He attended Chapman University's School Psychology in Orange, CA starting in August 2013. He graduated with his specialist degree in school psychology from Chapman University in May 2016 after completing his internship with Norwalk-La Mirada Unified School District and the Center for Autism and Neurodevelopmental Disorders. He transferred to the University of Missouri, Columbia for his doctoral degree in school psychology in August 2016. Jared will complete his pre-doctoral internship with Kansas City School Psychology Internship Consortium in August 2020.