University of Nevada, Reno

Mental Health Screeners in Elementary Schools:
Measurement Invariance Across Racial and Ethnic Groups

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Special Education and Disability Studies

by

Gregory D. Govan

Dr. Tammy V. Abernathy and Dr. Stephen L. Rock, Dissertation Advisors

December, 2017
We recommend that the dissertation prepared under our supervision by

**Gregory Dickinson Govan**

Entitled

**Mental Health Screeners in Elementary Schools: Measurement Invariance Across Racial and Ethnic Groups**

be accepted in partial fulfillment of the requirements for the degree of

**Doctor of Philosophy**

Tammy V. Abernathy, Ph.D., and Stephen L. Rock, Ph.D., Advisors

Cleborne Maddux, Ph.D., Committee Member

Shanon Taylor, Ed.D., Committee Member

Anthony Papa, Ph.D., Graduate School Representative

David W. Zeh, Ph. D., Dean, Graduate School

December, 2017
Abstract

Mental health screeners need to demonstrate measurement equivalence across the populations of their intended use in order to improve the fairness in the identification of students in need of social, emotional, and behavioral supports. This study examined measurement invariance on three mental screeners across five racial and ethnic groups. The *Elementary Social Behavior Assessment* measures academic enablers associated with the latent construct of teachability (ESBA). The *Student Risk Screening Scale* assesses externalizing problems (SRSS) and the *Student Internalizing Behavior Screener* measures internalizing problems (SIBS). Multigroup confirmatory factor analyses tested for measurement invariance from the sample of African American (18%), Asian American/Pacific Islander (13%), Latino Hispanic (25%), European American (31%), and multiracial (11%) groups of students in elementary schools. Only the ESBA required respecification to establish an adequate baseline model. The ESBA, SRSS, and SIBS demonstrated metric invariance with ordinal ratings of *never*, *occasionally*, *sometimes*, and *frequently* in addition to scalar invariance with the thresholds between the ordinal ratings. Thus, the total scores from the ESBA, SRSS, and SIBS generalize across racial and ethnic groups and the student’s race or ethnicity is less likely to mask their true level of need for social, emotional, and behavioral supports. The results indicate that the ESBA, SRSS, and SIBS may help teachers to identify racially and ethnically students who need intervention, to customize the interventions, and to evaluate students’ response to intervention. Schools using these mental health screeners may reduce disproportionality in discipline and special education.
Dedication

This research project is dedicated to teachers and students.

Challenging teachers create the conditions for learning.

Likewise, challenging students become our best teachers.

“Truly I tell you, whatever you did for one of the least of these brothers and sisters of mine, you did for me.” -Matthew 25:40
Acknowledgements

This dissertation would not have been possible without the acceptance, guidance, and support from Dr. Tammy Abernathy. You opened the door of opportunities for me to pursue my own learning path at UNR. Thank you Dr. Abernathy for your acceptance of someone as willful as me. Your support over the years helped me come around and learn academic and experiential lessons from doctoral studies. You model great teaching.

Dr. Stephen Rock leveled with me about the importance of respecting the culture of higher education in pursuit of goals beyond the university. Thank you for supporting my constant concern for family throughout the program.

Dr. Shanon Talyor, thank you for offering multiple opportunities to collaborate. Your incidental teaching provided meaningful learning.

Dr. Clebe Maddux, thank you for demonstrating ways to keep statistical analyses understandable so that a larger audience may interpret the results.

Dr. Tony Papa, thank you for joining my committee years after the class on multiple regression. Furthermore, thank you for instilling the curiosity to examine the meaning in variance from various sources. Your introduction to SEM made this research possible.

Kelcey Schmitz, MEd, you are an et al. angel. This entire dissertation rests upon your collaboration. Thank you for courage to speak truth to power while spreading the Good News of Positive Behavioral Interventions and Supports.

The inspiration for serving others through education comes from my family. My mom, Linda Govan, encouraged me to live the adventure and taught me to “breathe”
through challenges. My dad, Greg Govan, gave wise counsel and modeled compassionate care of others. I appreciate the encouragement and editing over the years by my sister, Carrie Govan Skelly, and her husband, Kevin.

Katy, you helped me pursue my dreams. Thank you for walking with me around the lake in order to get across it. Maya, your persistence in the face of challenges inspires me to become a better person. Katy and Maya, may we go on adventures! Also, now that the dining table is free of dissertation materials, let’s celebrate by hosting parties.

Ultimately, I extend much gratitude for all of my teachers and students.
# Table of Contents

Abstract................................................................................................................................. i

Dedication................................................................................................................................. ii

Acknowledgements................................................................................................................... iii

List of Tables ........................................................................................................................... xi

List of Figures .......................................................................................................................... xiii

Chapter 1: Introduction ............................................................................................................ 1
  Disproportionality in Special Education and School Discipline ........................................ 1
  Mental Disorders Among School-Aged Children ................................................................. 4
  National Recommendations on Early Intervention for Mental Health Problems ............. 6
  Traditional Approach to Identification of Students in Need of Additional Support .......... 7
  Early Identification is an Alternative to Wait-to-Fail Expectations .................................... 9
  Universal Screeners’ Function as an Aid to Decision Making ............................................ 9
  Measurement qualities of screeners. .................................................................................... 10
  Measurement error and bias ................................................................................................ 10
  Mental Health Screeners in this Study ............................................................................... 12
  Problem Statement .............................................................................................................. 13
  Purpose Statement .............................................................................................................. 14
  Research Questions ............................................................................................................ 15
  Research Question 1 ........................................................................................................... 16
    Question 1.1 .................................................................................................................... 16
    Question 1.2 .................................................................................................................... 16
    Question 1.3 .................................................................................................................... 16
  Research Question 2 ........................................................................................................... 16
    Question 2.1 .................................................................................................................... 17
Question 2.2 .................................................................................................................. 17
Question 2.3 .................................................................................................................. 17
Research Question 3 ........................................................................................................ 17
Question 3.1 .................................................................................................................. 17
Question 3.2 .................................................................................................................. 17
Question 3.3 .................................................................................................................. 18
Significance of the Study ............................................................................................... 18
Definition of Terms ........................................................................................................ 19
Organization of This Study ............................................................................................ 23
Chapter Two: Literature Review .................................................................................... 24
Disproportionality .......................................................................................................... 25
  Disproportionality in school discipline ................................................................. 26
  Disproportionality in special education ................................................................. 29
  Outcomes of disproportionality in school discipline and special education ........ 31
Internalizing and Externalizing Spectra of Mental Health Problems ......................... 33
  Prevalence of mental health disorders in school-age children ......................... 34
  Educational outcomes of externalizing and internalizing problems .................. 36
Disproportionate Unmet Mental Health Needs ............................................................ 37
National Recommendations on Universal Screening for Mental Health Problems .... 38
  Traditional teacher referral ..................................................................................... 39
  Tolerance theory of teachability ............................................................................. 39
  Universal screeners enhance professional judgment and decision making ........ 43
  Universal mental health screening procedure ..................................................... 44
  Psychometric properties of the scores from screeners ....................................... 45
  Validity and fairness ................................................................................................. 46
Constructs in this Study ................................................................. 47
Teachability .................................................................................. 48
Externalizing problems ................................................................. 49
Internalizing problems ................................................................. 49
Mental Health Screeners in this Study .......................................... 51
Elementary Social Behavior Assessment .................................... 52
Student Risk Screening Scale ..................................................... 53
Student Internalizing Behavior Screener .................................... 55
Multigroup Confirmatory Factor Analysis ................................... 57
Measurement Invariance .............................................................. 58
Configural invariance ................................................................... 58
Metric invariance ......................................................................... 59
Scalar invariance .......................................................................... 59
Summary ....................................................................................... 60

Chapter Three: Methods ............................................................... 62
Research Design ........................................................................... 62
Participants and Setting ............................................................... 63
Sampling method .......................................................................... 63
Selection of participating schools ............................................... 63
Selected schools .......................................................................... 65
Demographics of the teachers and administrators ....................... 69
Selection of student sample ......................................................... 70
Sample size .................................................................................. 72
Instrumentation ............................................................................. 72
Elementary Social Behavior Assessment .................................... 73
Student Risk Screening Scale ........................................................................................................... 74
Student Internalizing Behavior Screener............................................................................................ 75
Procedures for Data Collection ........................................................................................................ 76
Data Analysis ...................................................................................................................................... 77
Measured variables ............................................................................................................................ 78
Data analysis procedures ................................................................................................................ 78
Steps for testing for measurement invariance with confirmatory factor analysis ......................... 78
  Research question 1. ......................................................................................................................... 79
  Research question 2. ......................................................................................................................... 80
  Research question 3. ......................................................................................................................... 81
Evaluation of the goodness of fit for models. ...................................................................................... 82
Evaluation of measurement invariance in nested models. ................................................................. 83
Chapter Four: Results ....................................................................................................................... 85
Description of the Sample and Data ................................................................................................. 85
  Demographic data. ............................................................................................................................ 86
  Power .............................................................................................................................................. 86
  Missing data .................................................................................................................................... 87
  Data normality ............................................................................................................................... 87
Results ............................................................................................................................................... 89
  Elementary Social Behavior Assessment ...................................................................................... 89
  Student Risk Screening Scale ......................................................................................................... 95
  Student Internalizing Behavior Scale ............................................................................................. 97
Preliminary results to the research questions. .................................................................................... 99
  Research question 1. ......................................................................................................................... 99
  Research question 2. ......................................................................................................................... 99
List of Tables

Table 1 Inclusion Criteria for Selected Schools ..........................................................66
Table 2. Selected Schools and National Elementary School Racial and Ethnic Demographics ...........67
Table 3 Demographics and Educational Experience at the Selected Schools by Percentage ...............70
Table 4 Demographics of Students by Percentage ..................................................................71
Table 5 Univariate Skewness and Kurtosis ...........................................................................88
Table 6. Initial Unconstrained Model on the Elementary Social Behavior Assessment ................90
Table 7 Respecified Unconstrained Models with Covarying Residuals on the Elementary Social Behavior Assessment .................................................................93
Table 8 Unconstrained Models for the Student Risk Screening Scale ........................................96
Table 9 Unconstrained Models on the Student Internalizing Behavior Scale .............................98
Table 10 Tests of Measurement Invariance on Respecified ESBA with Covarying Residuals ..........102
Table 11 Tests of Measurement Invariance on Student Risk Screening Scale ............................103
Table 12 Tests of Measurement Invariance on the Student Internalizing Behavior Scale. ..............104
Table 13 Configural Invariance Estimated with WLSMV, WLSM, ULSM, and MLR .....................190
Table 14 Metric Invariance Estimated with WLSMV, WLSM, ULSM, and MLR ............................192
Table 15 Scalar Invariance Estimated with WLSMV, WLSM, ULSM, and MLR ............................194
Table 16 Polychoric Correlation Matrix for ESBA .....................................................................196
Table 17 Polychoric Correlation Matrix for SRSS ....................................................................196
Table 18 Polychoric Correlation Matrix for SIBS .....................................................................197
Table 19 Polychoric Residuals for the ESBA ............................................................................198
Table 20 Parameter Estimates for Scalar Invariance on the ESBA for African American Students. ....200
Table 21 Parameter Estimates for Scalar Invariance on the ESBA for Asian American Students. ......201
Table 22 Parameter Estimates for Scalar Invariance on the ESBA for Latino Hispanic Students. ......202
Table 23 Parameter Estimates for Scalar Invariance on the ESBA for Multiracial Students ..............203
Table 24 Parameter Estimates for Scalar Invariance on the ESBA for European American Students   ....204
Table 25 Parameter Estimates for Scalar Invariance on the SRSS for African American Students .......205
Table 26 Parameter Estimates for Scalar Invariance on the SRSS for Asian American Students ..........206
Table 27 Parameter Estimates for Scalar Invariance on the SRSS for Latino Hispanic Students ..........207
Table 28 Parameter Estimates for Scalar Invariance on the SRSS for Multiracial Students ...............208
Table 29 Parameter Estimates for Scalar Invariance on the SRSS for European American Students ....209
Table 30 Parameter Estimates for Scalar Invariance on the SIBS for African American Students ..210
Table 31 Parameter Estimates for Scalar Invariance on the SIBS for Asian American Students ............211
Table 32 Parameter Estimates for Scalar Invariance on the SIBS for Latino Hispanic Students .......212
Table 33 Parameter Estimates for Scalar Invariance on the SIBS for Multiracial Students ..............213
Table 34 Parameter Estimates for Scalar Invariance on the SIBS for European American Students .......214
List of Figures

Figure 1. Risk ratio of office discipline referral from the elementary classroom.................................27
Figure 2. Risk ratio for eligibility for special education services..........................................................30
Figure 3. School sample and national population participation in selected programs..............................68
Figure 4. School sample and national population interrupted participation in instruction..........................69
Figure 5. Unstandardized Parameter Estimates for the respecified omnibus ESBA.................................94
Figure 6. Unstandardized Parameter Estimates for the SRSS...............................................................95
Figure 7. Unstandardized parameter estimates for the SIBS.............................................................98
Figure 8. Network of polychoric residuals..........................................................................................199
Figure 9. Scalar model of the ESBA for African American students....................................................200
Figure 10. Scalar model of the ESBA for Asian American students.....................................................201
Figure 11. Scalar model of the ESBA for Latino Hispanic students......................................................202
Figure 12. Scalar model of the ESBA for multiracial students.............................................................203
Figure 13. Scalar model of the ESBA for European American students..............................................204
Figure 14. Scalar model of the SRSS for African American students..................................................205
Figure 15. Scalar model of the SRSS for Asian American students....................................................206
Figure 16. Scalar model of the SRSS for Latino Hispanic students.....................................................207
Figure 17. Scalar model of the SRSS for multiracial students.............................................................208
Figure 18. Scalar model of the SRSS for European American students..............................................209
Figure 19. Scalar model of the SIBS for African American students...................................................210
Figure 20. Scalar model of the SIBS for Asian American students.....................................................211
Figure 21. Scalar model of the SIBS for Latino Hispanic students.....................................................212
Figure 22. Scalar model of the SIBS for multiracial students.............................................................213
Figure 23. Scalar model of the SIBS for European American students..............................................214
Chapter 1: Introduction

As the school-age population becomes more racially and ethnically diverse, fulfilling the promise of public education requires increased vigilance for equity and fairness for all students. The traditional practice of teacher identification of students in need of assistance coincides with problems of disproportionality and school discipline in special education. Universal mental health screening adds an evidence-based alternative to relying only on subjective judgments. The statistical enhancement for identifying students in need of additional support promises to work fairly for African American, Asian American, Latino Hispanic, European American and multiracial students.

According to the Code of Fair Testing Practices in Education from the American Psychological Association (APA), fairness refers to assessment practices that provide reliable and valid information regardless of student characteristics (e.g., racial and ethnic background) unrelated to the purpose of the measurement instrument (Joint Committee on Testing Practices, 2004). In light of increasing racial and ethnic diversity, this study examines the fairness of measurement properties of the mental health screeners. Specifically, the study evaluates whether the latent constructs of internalizing problems, externalizing problems, and teachability share the same meaning across African American, Asian American, Latino Hispanic, and multiracial groups compared to European American students in elementary schools.

Disproportionality in Special Education and School Discipline

The long history of inequality in education for racial and ethnic minorities continues to this day, revealed in the disproportionate representation of racial and ethnic
groups in educational practices that remove students from the general education classroom. Disproportionality arises when racial or ethnic background influences the probability of referral for disciplinary or special education. Disproportionality refers to the over-representation or under-representation of specific racial and ethnic groups in special education for disabilities relative to their proportion in overall enrollment in schools.

Nearly 50 years ago, the president of the Council for Exceptional Children called into question the practice of placing a higher proportion of students from racial and ethnic groups into special education than were currently reflected in the general population of the nation (Dunn, 1968). According to National Academy of Sciences (NAS), African American, American Indian, and Native Alaskan students were over-represented in nearly every category of special education, and Latino Hispanic and Asian American students were under-represented in special education (Donovan & Cross, 2002). Furthermore, the Office of Special Education and Rehabilitative Services (OSERS) reported on continued racial and ethnic disparities in special education through 2015 (US Department of Education, OSERS, 2016b).

Furthermore, marked disproportionality appears in the more subjectively defined special education category of emotional-behavioral disability (EBD). Eligibility for this category relies more heavily on the professional judgment of the evaluation team compared to other categories such as hearing or vision impairment (Donovan & Cross, 2002). By contrast, the determination of vision impairment relies on the measurement of the degree of blindness or partial sight—measurements far more specific and concrete. The subjective evaluation of challenging behavior may contribute to disproportionality in
special education (Zhang & Katsiyannis, 2002). Currently, compared to the index of White students in special education under the EBD category, African American students receive services under the EBD category at rate 39% higher, multiracial students at 29% higher, Latino Hispanic students receive them at rate 35% lower, and Asian American students at 62% lower rate (National Center for Education Statistics, 2015). The high variability in the likelihood of eligibility for special education under the EBD category suggests that both under-representation and over-representation occur based on race and ethnicity.

Disproportionality also arises in school discipline practices for racially and ethnically diverse students. The subjective interpretations of student behaviors may influence the perceived need for school discipline. For example, disrespect and disruption occur along a continuum from minor to major degrees of violations of teacher expectations. Whereas one teacher may address minor rule violations in class, another teacher frequently may send students out of class to the school office for major offenses, with the expectation that administrators will apply punitive consequences. Office discipline referrals for disruptive and disrespectful behaviors top the list of reasons for all students leaving the classroom (Skiba et al., 2014; Skiba et al., 2011).

The consequences for office discipline referrals are subject to disproportionality as well (Rocque, 2010). A disproportionate percentage of African American students experience more frequent disciplinary procedures and a higher intensity of disciplinary consequences compared to other racial and ethnic groups (Bradshaw, Mitchell, O’Brennan, & Leaf, 2010). The Civil Rights Project report on national trends in disciplinary practices in elementary schools found that, while 2.6% of all elementary
students were suspended in 2011-2012 academic year, 7.6% of African American students, 2.9% of Native Americans, 2.1% of Latino Hispanic, 1.6% White, 1.2% Pacific Islander, and 0.5% of Asian American students were suspended at least once (Losen, Hodson, Keith II, Morrison, & Belway, 2015).

**Mental Disorders Among School-Aged Children**

Student behaviors that may contribute to eligibility for special education or for school discipline practices often are associated with mental health concerns (Bogart et al., 2013). In the case of impulsivity or aggression, the teacher typically makes a referral for discipline. When discipline does not immediately improve the behavior, the teacher may then refer the student for evaluation for special education. Similarly, teachers may struggle initially to handle more perplexing behaviors such as temper tantrums, inappropriate crying, social withdrawal, and refusal to participate in activities. However, once these moody or anxious behaviors become unmanageable, the teacher often makes a referral to special education evaluation. Students suffering from mental health problems also exhibit these behaviors. Thus, the problem of disproportionality must be examined within the larger context of mental health.

Based on the prevalence rates, students will encounter mental health problems directly through personal experience or indirectly through a peer for some period during their educational careers. Over the course of a lifetime, the cumulative prevalence for mental disorders reaches 49.5% of children and youth by the time they reach the age of eighteen (Merikangas et al., 2010). Even within any given year, the Centers for Disease Control and Prevention (CDC) estimates annual prevalence between 13 and 20% of children and youth between ages five and seventeen experience a mental disorder (Perou
et al., 2013). These rates indicate that most public school classrooms may expect to include between two and five students with a mental disorder (Forness, Freeman, Paparella, Kauffman, & Walker, 2012; Forness, Kim, & Walker, 2012).

A mental disorder meets the full criteria of the Diagnostic and Statistical Manual of Mental Disorders (DSM; American Psychiatric Association, 2013). However, mental health problems that do not meet the full criteria still exert substantial functional impairment and distress (Lewinsohn, Shankman, Gau, & Klein, 2004) and are referred to as minor, sub-syndromal, or sub-clinical disorders (Bertha & Balázs, 2013). At the sub-clinical level, the semi-annual prevalence of mental health problems affects 23.1% of children between ages six and eleven (Simon, Pastor, Reuben, Huang, & Goldstrom, 2015). These differing prevalence rates highlight the greater prevalence of minor, sub-threshold mental health problems relative to more severe, clinical mental disorders.

Mental health problems appear along broad dimensions of internalizing and externalizing problems (Brown & Barlow, 2005; Widiger & Samuel, 2005). Internalizing mental problems refer to the broad inward expression of anxious, depressive, withdrawn, and somatic complaints (Ask, Waaktaar, Seglem, & Torgersen, 2016). The spectrum of internalizing problems encompasses the co-occurrence of anxiety and mood disorders (Mclaughlin & King, 2015). Externalizing mental health problems refer to the broad outward exhibition of impulsivity, irritability, aggression, and defiance of social norms (Olson et al., 2013; Walton, Ormel, & Krueger, 2011). The spectrum of externalizing problems captures combinations of attention deficit/hyperactivity disorder, oppositional defiant disorder, and conduct disorder (Beauchaine & McNulty, 2013; Tackett, 2010). Within either internalizing or externalizing spectra, the presentation of symptoms may
include subclinical levels from several disorders while never reaching the threshold for any specific categorical diagnosis (Blanco et al., 2015; Carragher, Krueger, Eaton, & Slade, 2015; Coghill, & Sonuga-Burke, 2012; Eaton, 2015).

**National Recommendations on Early Intervention for Mental Health Problems**

In recognition of the adverse effects of mental health problems in the lives of children and youth, multiple policy makers and government agencies have promoted universal screening in schools for early identification and prevention. For instance, the Institute of Medicine recommends mental health screening in education systems (O'Connell, Boat, & Warner, 2009). Currently, the re-authorization of the Every Student Succeeds Act (ESSA) recommends mental health screening as a component of multi-tiered systems of support in activities to support safe and healthy students (Every Student Succeeds Act of 2015, Pub. L. No. 114-95 § Section 4108 [2015-2016]). The ESSA takes effect in 2017-2018. Multi-tiered systems of support, as a service delivery model, emphasize prevention and early intervention as alternatives to the traditional wait-to-fail model of identification before intervention (Brown-Chidsey & Bickford, 2016).

Prevention proactively screens children to identify needs rather than waiting for students to fail (Donovan & Cross, 2002; Levitt, Saka, Romanelli, & Hoagwood, 2007). In contrast to the wait-to-fail model that emphasizes deficits in the student, the prevention approach seeks to identify problems in the contextual fit between student needs and the learning context that addresses those needs (Kettler, Glover, Albers, & Feeney-Kettler, 2014). Prevention closes the gap between the identification of student needs and early interventions in the learning context by addressing minor problems before they develop into severe or chronic diagnosable disorders (O'Connell et al., 2009).
Traditional Approach to Identification of Students in Need of Additional Support

The traditional approach to identifying students who need additional social, emotional, and behavioral support primarily relies upon the judgment of teachers. In such cases, the teacher requests assistance from either the child study team or the school administration to address the behavior of the student. These educators directly support the teacher and, ideally, the student as well. Most often, the immediate response to the teacher referral for assistance provides indirect support to the student.

In a typical classroom, teachers manage the complex needs of the whole class with a finite range of instructional resources of knowledge, skills, relationship, time, and effort (Brown & Saks, 1987). The teacher constantly adjusts the allocation of instructional resources between the learning needs of the majority of the class and the individual needs of a few students of concern (Gerber & Semmel, 1984). The teacher deems students as teachable based on their positive response to the instructional resources. The push and pull from the dynamic interaction among the teacher’s personal instructional resources, the learning needs of the majority of the class, and the individual needs of some students create the context for problems to arise (Gerber & Semmel, 1985). In the case of inadequate response to instructional resources, the student’s needs surpass the teacher’s threshold of tolerance for teachability.

The subjective boundaries of teachability shift according to the teacher’s tolerance for behavioral diversity. For example, in one classroom, an animated student may appear hyperactive relative to calmer peers because the active student stands out compared to the rest of the class. In this situation, the teacher notices the difference in the amount of instructional effort when the active student is absent compared to when the same student
is present. The presence of the active student places greater demands on finite instructional resources. In absence of the active student, the teacher notices the majority of the class profits more from the instructional resources. Understandably, the teacher may believe that the active student interferes with the learning of the class even though all of the students are making academic progress. However, if this student attended another classroom containing energetic learners, the teacher might not notice a significant change in the instructional effort depending on the attendance of the active student. The mismatch between teacher expectations and student performance can depend on the context rather than a problem inherent in the student.

The subjective boundaries of teacher tolerance for behavioral diversity can lead to bias in the identification of students in need of additional support (Shinn, Tindal, & Spira, 1987). In the traditional model of teacher referrals, biases occur to varying degrees of tolerance for specific student behaviors or student characteristics such as gender, racial, or ethnic background (Lane, Pierson, & Givner, 2003; Lane, Wehby, & Cooley, 2006; Walker, 1985). A meta-analysis by Tenenbaum and Ruck (2007) showed that elementary teachers refer more students from diverse racial and ethnic groups for discipline or special education compared to European American students ($d = .46$). When referral rates differ from population rates for racial and ethnic groups at such a level, the validity of subjective judgment comes into question.

The subjective judgment inherent in the traditional model of teacher referral misses identifying students who need social, emotional, and behavioral supports compared to a more systematic and structured approach using a universal screening procedure (Kamphaus, Distefano, Dowdy, Eklund, & Dunn, 2010). The traditional
approach demonstrates lower accuracy for identifying students both needing and not needing additional support (Eklund et al., 2009; VanDerHeyden & Witt, 2005). This lower accuracy allows students to fail while delaying opportunities to intervene earlier (Eklund & Dowdy, 2014).

**Early Identification is an Alternative to Wait-to-Fail Expectations**

The alternative public health approach to the traditional teacher referral for assistance emphasizes screening, prevention, and intervention to address mental health concerns, especially for internalizing and externalizing problems (Gutkin, 2012; O'Shaughnessy, Lane, Gresham, & Beebe-Frankenberger, 2003; Stiffman et al., 2010). Universal screening assists teachers by detecting milder levels of need within their threshold of tolerance, especially for internalizing behaviors which often are overlooked (Weist, Rubin, Moore, Adelsheim, & Wrobel, 2007). Prevention improves opportunities for healthy emotional, social, and educational development while ethically reducing suffering (Cruden, Kelleher, Kellam, & Brown, 2016; Greenberg, Domitrovich, & Bumbarger, 2001). Universal mental health screening practices hold promise for early identification of students in need of additional social, emotional, and behavioral supports (Dowdy et al., 2015).

**Universal Screeners’ Function as an Aid to Decision Making**

In order to serve as an aid to decision-making, screeners must improve the accuracy in identifying students in need of additional support compared to the currently subjective approach of teacher referral. In universal screening procedures, teacher judgment continues to play an important role in making decisions based upon the subjective integration of direct observations and interactions with students (Walker et al.,
2015). However, adding the statistical data from universal screening to the traditional subjective model provides structure to professional judgment via evidence-based rules for interpreting the scores from measurement instruments concerning the student (Falzer, 2013).

**Measurement qualities of screeners.** Making valid decisions based on screening instruments depends on the psychometric properties of reliability and validity. Reliability indicates the degree to which the scores from a screening instrument consistently measure the same construct. On one hand, if the screening scores are unreliable, then it is difficult to determine whether the scores truly represent the measured construct. On the other hand, when scores are consistently precise, then it is possible to consider the meaning of the scores. The validity of the decisions based upon screeners relies on both construct and predictive validity. Construct validity refers to the degree to which the instrument measures the construct that it intends to measure (Cronbach & Meehl, 1955). Screening instruments measure latent constructs that are not directly observable, such as internalizing or externalizing problems and the construct of teachability. The foundation of predictive validity stands upon the screener’s ability to correctly classify students on the measured construct (Millsap, 2007).

**Measurement error and bias.** The scores on any assessment will always contain some measurement error. In the case of systematic error, the factors that affect the precision of measurement produce consistently upward or downward bias. These sources of variance from unrelated factors increase the variance in a score from influences other than the intended latent construct.
Metric variance and scalar variance exemplify systematic error. Metric variance refers to variance around the precision of the unit of measurement. Many screeners rank the frequency of behaviors with the ordinal descriptors of never, occasionally, sometimes, or frequently. However, such descriptors may not share the same meaning for the construct of interest when applied to students from each racial and ethnic group. For example, the ratings of occasionally and sometimes may vary depending on the ethnic group to which the student belongs. Scalar variance refers to variance around the origin point of measurement. Systematic error occurs when the initial value of a scale consistently starts with different quantities based on factors unrelated to construct of interest.

These metric and scalar sources of error in the scores add variance that contributes to either over- or under-estimation of a student’s level of the construct. When a student’s membership in a racial or ethnic group influences the measurement of the latent construct of interest, the systematic error introduces bias into the measurement. Bias from measurement error reduces the accuracy in classification of students in need of additional support and, thereby, does not predict valued outcomes for those students. Systematic measurement error due to racial and ethnic background compromises the validity of the screener.

In order for screeners to improve upon teacher judgment, they need to demonstrate that they are valid for all students. The validity of the latent constructs of internalizing and externalizing problems and teachability depends on establishing the equivalent meaning and measurement properties for each ethnic group within the sample so that the results may be generalized across represented populations (Millsap, 2011).
When a screener fails to establish measurement invariance, practitioners and researchers may interpret spurious differences among racial and ethnic groups that really may be artifacts of measurement (Meredith & Teresi, 2006). Conversely, measurement artifacts may mask true differences among groups and practitioners, and researchers may also fail to identify them. In the case of screening diverse racial and ethnic groups for selection for intervention, either truly healthy students may be falsely identified as needing additional supports or other students truly in need of support may fall through the cracks (Chen, 2008). Reducing measurement bias in universal screeners supports teachers making valid decisions for diverse racial and ethnic groups.

**Mental Health Screeners in this Study**

This study extends the research on three freely available mental health screeners. Teachers rate each student on separate screening instruments according to their perceived frequency of the indicated behaviors. The range of frequency extends from *never, occasionally, sometimes*, and *frequently*. The Elementary Social Behavior Assessment (ESBA; Pennefather & Smolkowski, 2015) measures the frequency of academic enabling behaviors that support teachability (DiPerna & Elliot, 2000; Elliot, DiPerna, Mroch, & Lang, 2004). The Student Risk Screening Scale (SRSS: Drummond, 1994) measures the frequency of violations of social norms associated with externalizing behavior problems. The Student Internalizing Behavior Screener (SIBS; Cook et al., 2011) measures the occurrence of negative emotions and social withdrawal associated with internalizing behavior problems. Although the three mental health screeners have demonstrated feasibility, reliability, and validity, none of the screeners has addressed measurement invariance across racial and ethnic groups.
Problem Statement

As racial and ethnic diversity increases in the composition of elementary schools across the country, so does the need to empirically evaluate the equity and fairness of educational practices. In particular, universal mental health screening may reduce disproportionality in discipline and special education services for students from diverse racial and ethnic backgrounds. Screening all students in a school serves the purpose of early identification and intervention through the accurate classification of mental health needs and the prediction of future outcomes. In order to assist teachers with making fair decisions regarding African American, Asian American, Latino Hispanic, European American, and multiracial students in elementary schools, universal screening instruments need to establish measurement properties that support valid interpretations.

Making fair decisions with universal screening instruments relies on adequate psychometric properties. As is well known, the ethnic composition of the sample from which the screener was designed may affect the reliability. If the reliability of the screening instruments varies from one ethnic group to another, the validity of the inferences from this screener comes into question. In such circumstances, the race and ethnicity of the person being measured introduces bias that pushes scores upward or downward based on factors unrelated to the construct of interest. Measurement bias in universal screeners attenuates the assistance to making fair decisions. Therefore, in order to enhance the professional judgment of teachers, universal screeners should demonstrate measurement invariance for each of the racial and ethnic groups under consideration.

The design of screening instruments should minimize construct-irrelevant variance that interferes with fair interpretations according to the *Standards for*
Educational and Psychological Testing set by the Joint Committee of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education in 2014. Accordingly, the constructs of internalizing and externalizing problems and the construct of teachability should share the equivalent meaningful structure for each factor across each racial and ethnic group in the sample. Furthermore, the rating descriptors of never, occasionally, sometimes, and frequently should measure the equivalent metric unit across the racial and ethnic groups. Lastly, the meaning of never, occasionally, sometimes, and frequently on the scale should measure from the equivalent origin across the groups. These three levels of equivalence demonstrate measurement invariance. None of the cost-free, universal screeners have demonstrated measurement equivalence among five ethnic groups in the same sample.

This study addresses the gap in the research on the development of universal screeners for fair interpretation across five racial and ethnic groups. Specifically, this study examines the problems of the equivalent meaning and the equivalent measurement for each of the latent constructs in three universal screening instruments based on a sample of African American, Asian American, Latino Hispanic, European American, and multiracial students in elementary schools. The three screening instruments include the Elementary Social Behavior Assessment (Pennefather & Smolkowski, 2015), the Student Risk Screening Scale (Drummond, 1994), and the Student Internalizing Behavior Screener (Cook et al., 2011).

**Purpose Statement**

Universal screeners promise to enhance teacher judgment regarding the identification of students with mental health needs. Measurement invariance across racial
and ethnic groups supports the fair interpretation of the universal screeners. The purpose of this study is to test for measurement invariance on the three universal screeners at three nested levels involving groups of African Americans, Asian American-Pacific Islanders, Latino Hispanic, European Americans, and multi-race students in elementary schools. Initially, each screener should measure the same construct of interest for each ethnic group. The conceptual equivalence for each latent construct requires the same meaningful, or nontrivial, factor structure for each ethnic group. The screener must measure the same construct for each racial and ethnic group as a condition for the interpretability of the obtained scores. The next level of measurement invariance examines the degree to which the rating descriptors of never, occasionally, sometimes, and frequently measure the equivalent metric unit of the particular construct across the racial and ethnic groups. The last level of measurement invariance examines the origin point for the scale. Such screening instruments with scores that hold measurement equivalence across ethnic groups assist teachers with improved fairness for identifying students in need of mental health support.

**Research Questions**

In order to address the problem of the fair interpretation of universal screeners across five racial and ethnic groups, this study proposed to examine measurement invariance for each instrument. The overarching question for each screening instrument is: Does the screener demonstrate measurement equivalence for each group of African American, Asian-Pacific Islander, Latino Hispanic, and multiracial students compared with the index group of European Americans students in elementary schools? This larger question tested measurement invariance at three nested levels in which each successive
level depends upon the previous one. The tests for each of the nested levels of
measurement invariance examined one screening instrument at a time.

**Research Question 1.** The first level of research questions for each screener tests
whether the screeners measure the same construct of interest across each ethnic group by
examining the factors loading from each indicator for the latent variable. Does the
screener measure the latent construct of interest with the equivalent factor structure
consisting of nontrivial indicators across each of the racial and ethnic groups relative to
the European American index group?

**Question 1.1.** Does the Elementary Social Behavior Assessment measure the
latent construct of teachability with the equivalent factor structure across each of the
racial and ethnic groups?

**Question 1.2.** Does the Student Risk Screening Scale measure the construct of
externalizing problems with the equivalent factor structure each of the racial and ethnic
groups?

**Question 1.3.** Does the Student Internalizing Behavior Screener measure the
construct of internalizing problems with the equivalent factor structure across each of the
racial and ethnic groups?

**Research Question 2.** The second level of research questions tests whether
the latent construct shares the same metric unit of frequency across racial and ethnic
groups by examining the slope of the factor loading while fixing the indicators to the
latent variable. Do the ordinal descriptors of *never, occasionally, sometimes,* and
*frequently* demonstrate measurement equivalence in the metric unit of frequency on for
latent construct of interest for the screener across each of the racial and ethnic groups relative to the European American index group?

**Question 2.1.** Do the ordinal descriptors demonstrate measurement equivalence in the metric unit for latent construct of teachability on the Elementary Social Behavior Assessment across each of the racial and ethnic groups?

**Question 2.2.** Do the ordinal descriptors demonstrate measurement equivalence in the metric unit for latent construct of externalizing problems on the Student Risk Screening Scale across each of the racial and ethnic groups?

**Question 2.3.** Do the ordinal descriptors demonstrate measurement equivalence in the metric unit for latent construct of internalizing problems on the Student Internalizing Behavior Screener across each of the racial and ethnic groups?

**Research Question 3.** The third level of research questions addresses whether the scale holds the same origin on the scale across each ethnic group by examining the threshold of the starting value while fixing the metric unit constant. Does the descriptor, *never*, establish measurement equivalence on the initial threshold of the scale of the construct of interest for the screener across each of the racial and ethnic groups relative to the European American index group?

**Question 3.1.** Does the descriptor, *never*, establish measurement equivalence on the initial threshold of the scale of teachability on the Elementary Social Behavior Assessment across each of racial and ethnic groups?

**Question 3.2.** Does the descriptor, *never*, establish measurement equivalence on the initial threshold of the scale of externalizing problems on the Student Risk Screening Scale across each of racial and ethnic groups?
Question 3.3. Does the descriptor, *never*, establish measurement equivalence on the initial threshold of the scale of internalizing problems on the Student Internalizing Behavior Screener across each of racial and ethnic groups?

Significance of the Study

Universal screeners that demonstrate measurement invariance hold significant implications for schools, teachers, and students. Students may benefit from early identification and intervention. Universal screening presents an opportunity to intervene early (Brown & Barlow, 2005). Students in need of social, emotional, and behavioral supports will likely receive assistance sooner than the traditional model. Screeners provide a basic measurement of different levels of functioning across internalizing and externalizing problems in addition to prosocial skills. Understanding a student’s level in these various dimensions enables teachers and interventionists to tailor the supports according to strengths and concerns (O’Connell et al., 2009; Walker & Gresham, 2015). These targeted supports may better address social, emotional, and behavioral needs than generic interventions. Screening may ensure that students have an equal opportunity to have their social, emotional, and behavioral needs met while potentially addressing problems of disproportionality (Vincent, Randall, Cartledge, Tobin, & Swain-Bradway, 2011). Early detection of social, emotional, and behavioral concerns promises the opportunity to address the mental health needs of students before they surpass teacher tolerance for behavioral diversity or develop into chronic and severe impairment.

Teachers may benefit from statistically enhanced decision making. The classroom data may help teachers focus their instructional effort toward equity and fairness (Raines, Dever, Kamphaus, & Roach, 2012). Teachers may implement universal
prevention practices in the classroom. The data may help teachers reach out to students in need of mental health supports rather than referring students for discipline or overlooking internalizing problems.

The results of screening data assist teachers, professional learning communities, and schools to identify needs, allocate resources, and measure the effects of prevention and interventions (Oakes, Lane, Cox, & Messenger, 2014). Specifically, the data from universal screening may answer which students respond well to the school-wide systems of support. The school leadership team may provide teachers with adequate support to address classrooms with a substantial quantity of students with high needs. The services based on screening may reach a greater number of students for prevention and intervention than otherwise would occur through the traditional model of wait-to-fail prior to identification (Merrell & Buchanan, 2006). The reduction in disproportionality may come from the increase in providing targeted social, emotional, and behavioral supports.

**Definition of Terms**

*Bias* refers to the systematic error in a test score and violates measurement invariance (Millsap, 2011). Additionally, the Standards for Educational and Psychological Testing state that bias arises from “construct-irrelevant components of test scores that differentially affect the performance of different groups of test takers and consequently the reliability/precision and validity of interpretations and uses of the test scores” (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, p.216).
Configural invariance or form equivalence refers to the pattern of free and fixed model parameters (e.g., indicators) appearing similar, if not identical, for each of the groups compared. In the configural model, the factor loadings and thresholds are free to vary while only the referent indicator is fixed.

Diagnostic and Statistical Manual of Mental Disorders (DSM) is the handbook published by the American Psychiatric Association that outlines the categories of mental disorders and the criteria for diagnosing them (American Psychiatric Association, 2013).

Disproportionality refers to the level of representation of a group that unexpectedly differs from other groups within a category. Racial and ethnic disproportionality in special education or school discipline occurs when the level of students from specific racial and ethnic groups appears significantly higher or lower compared to the general population or compared to the majority of the population.

Ethnicity defines groups of people who share common sociological, cultural, historical, ancestral, or linguistic similarities.

Externalizing disorders primarily involve conduct disorder and oppositional disorder as described in the DSM.

Externalizing problems comprise the class of behaviors that involve “conflict with other people and with social mores, such as fighting, lying, stealing” (Achenbach & Rescorla, 2007, p.15). The Student Risk Screening Scale measures the level of risk for externalizing problems.

A factor represents the construct of an unobserved or a latent variable (Price, 2017).

A factor loading describes “the correlation between each variable (e.g., a test item
or total test score) in the factor” (Price, 2017, p. 139).

A *factor threshold* is a model parameter with ordered categorical variables that approximates the intercept of a factor loading on an indicator.

*Fairness* refers to the validity of test score interpretations for intended uses for individuals from all relevant subgroups. A test that is fair minimizes the construct-irrelevant variance associated with individual characteristics and testing contexts that otherwise would compromise the validity of scores for some individuals. (AERA, APA, & NCME, 2014, p. 219)

*Internalizing disorders* present as clusters of comorbid disorders “characterized by depressed mood, anxiety, and related physiological and cognitive symptoms” (American Psychiatric Association, 2013, p. 13).

*Internalizing problems* “occur primarily within the self, such as depression, anxiety and somatic problems” (Achenbach & Rescorla, 2007, p.150). The *Student Internalizing Behavior Screener* measures the level of risk for internalizing problems.

*Measurement equivalence* refers to measurement invariance.

*Measurement invariance* establishes that the measurement properties of the test are equivalent across each of the groups in the population (Brown, 2015).

*Mental disorder* is a condition that is “characterized by alterations in thinking, mood, or behavior (or some combination) associated with distress and/or impaired functioning” (Office of the Surgeon General, 1999, p. 5).

*Mental health* “is characterized by the achievement of development and emotional milestones, healthy social development, and effective coping skills, such that mentally
healthy children have a positive quality of life and can function well at home, in school, and in their communities” (Perou et al., 2013, p. 3).

*Metric invariance* occurs when the factor loadings are equivalent across each of the groups in the comparison. In the metric model, the factor loadings are constrained across groups while the factor thresholds remain free to vary.

*Pattern invariance* refers to configural invariance.

*Race* arises from the social construct regarding the meaning of physical features such as hair, eye, and skin color.

*Scalar invariance* occurs when each of the factor thresholds and the factor loadings are equivalent across all of the groups. In the scalar model, the factor loadings and the factor thresholds are constrained across groups.

A *screening test* is a type of test used “to make broad categorizations of test takers as a first step in selection decisions or diagnostic process” (AERA et al., 2014, p.223).

*Social, emotional, and behavioral problems* describe “difficulties that maybe early signs or symptoms of mental disorders but are not frequent or severe enough to meet the criteria for a diagnosis” (O'Connell et al., 2009, p. xxvi).

*Teachability* describes the latent construct of a class of behaviors that refer to the skills students need perform in order to benefit from and participate in instruction (Gerber & Semmel, 1984). The *Elementary Social Behavior Assessment* measures the level of the latent construct of teachability.

*Validity* involves “a judgment or statistical estimate based on the accumulated evidence of how well scores on the test or instrument measure what they are supposed to measure” (Price, 2017, p. 104).
Organization of This Study

This research study involves five chapters. Chapter 1 includes the background of the study, statement of the problem, purpose of the study, research questions, significance of the study, definition of terms, and limitations of the study. The review of literature in Chapter 2 begins with the context of disproportionality by racial and ethnic groups in special education and school discipline along with the prevalence rates of mental disorders by racial and ethnic groups. Furthermore, Chapter 2 elaborates on the construct of teachability within the theory of teacher tolerance for behavioral diversity. The literature review closes on the universal mental screening as an alternative to traditional teacher referrals along with a thorough description of three mental health screening instruments. Chapter 3 presents the methodology for this study of multiple racial and ethnic groups by specifying the selection of participants, the instrumentation and latent variables, data collection, and procedures for data analysis. Chapter 4 will cover the study’s results. These findings will include demographic information on the students and teachers in addition to the tests of the research questions regarding measurement invariance. The summary of this study in Chapter 5 will discuss the findings relevant to practitioners and researchers, recommendations for future research, and conclusions.
Chapter Two: Literature Review

According to the Digest of Educational Statistics, the student body entering schools across the country is becoming increasingly diverse (Snyder, de Bray, & Dillow, 2016). In particular, European American students no longer comprise a majority as the proportion of students from other racial and ethnic backgrounds rises (Snyder et al., 2016, Table 203.50). This change in the racial and ethnic composition of classrooms prompts the need to evaluate education practices for biases (Turner, 2015). This review of the literature regarding psychometric bias in universal mental health screeners situates the function of the screeners within teacher decision-making in light of issues related to the unmet mental health needs of students and the disproportionality in school discipline and special education.

While diversity is increasing, the persistent over- and under-representation of diverse racial and ethnic students involved in school discipline or special education services continues to challenge educators (Civil Rights Data Collection, 2016; Dunn, 1968). The pattern of disproportionality in discipline and special education suggests a need to change the traditional process of teacher referral. In the traditional model, the teacher refers students to school discipline or special education evaluation when the student’s behavior surpasses the individual teacher’s tolerance (Gerber & Semmel, 1984). The subjective threshold of tolerance for behavioral diversity makes this form of decision-making vulnerable to bias (McIntosh, Girvan, Horner, & Smolkowski, 2014).

As an alternative to the traditional model, universal mental health screening promises to identify students in need of additional support before their behavior becomes intolerable. Universal mental health screening adds a statistical aid for structured
professional judgment. However, in order to enhance the teacher’s professional judgment, universal mental health screeners should demonstrate psychometric properties that enable fair decisions. Specifically, universal mental health screeners should provide evidence of measurement equivalence across the racial and ethnic groups. *Measurement equivalence* empirically tests the degree of bias through the method of multigroup confirmatory factor analysis of measurement invariance. This literature review examines in detail the most salient extant research surrounding this topic.

**Disproportionality**

Disproportionality refers to the under- or over-representation of students from diverse cultural backgrounds in school discipline and special education (Skiba, 2013). Establishing the risk ratio is a common method for determining the level of disproportionality by comparing the risk of receiving disciplinary action or special education services relative to a comparison group, namely European American students (Bollmer, Bethel, Garrison-Mogren, & Brauen, 2007; Hosp & Reschly, 2003). At the time when the Department of Education compiled the data related to disciplinary actions and special education services, European American comprised the majority of students in public schools. Therefore, the risk ratio provides an indicator of the likelihood of receiving disciplinary actions or special education services compared to European American students (Coutinho & Oswald, 2000).

The risk ratio helps to identify disproportionate practices much as the Equal Employment Opportunity Commission (EEOC) helps with the enforcement of the Civil Rights Act of 1964 (EEOC, 1978). Diverse racial and ethnic groups align with the majority group of European American students when the risk ratio is 1.00. Conversely,
diverse groups with a risk ratio above 1.00 are more likely and groups below 1.00 are less likely to receive disciplinary actions or special education services. Disproportionate practices raise red flags when the risk ratio exceeds the four-fifths rule. Specifically, the four-fifths rule indicates disproportionality when the risk ratio, measuring the degree to which risk for racial and ethnic groups differs from that for European American, exceeds or falls below four-fifths (> 1.25 or < 0.80).

**Disproportionality in school discipline.** Students from ethnically and racially diverse backgrounds receive disproportionately frequent disciplinary actions compared to European American students (Boneshefski & Runge, 2014). Disproportionality in the forms of differential selection for and treatment of discipline based on race and ethnicity has raised concern for many decades (Children’s Defense Fund, 1975). School discipline occurs at two levels: initially at the teacher level with the decision to document a discipline procedure with an office referral, and subsequently at the administrative level with the delivery of consequences. This review examines discipline at the teacher level.

Problem behaviors that often result in disciplinary action range from minor to major violations of norms and rules. Gion, McIntosh, and Horner (2014) found that the most common forms of minor behavior problems include defiance/disrespect (36%), disruption (22%), and physical aggression (19%) while common major behavior problems include defiance or disrespect (27%), physical aggression (27%), and disruption (21%). The teacher’s individual judgment regarding the intensity of the violation distinguishes between minor versus major behavior problems. For example, one teacher may perceive a given episode of disrespect and disruption a minor problem, while
another might perceive it as a moderate. Whereas the teacher normally handles minor problems in class, major problems warrant sending the student to the administration for discipline.

Disproportionality occurs at the point of subjective decision-making in regard to the application of disciplinary practices for minor infractions. Data from 436 schools in the School-wide Information System (SWIS; May et al., 2010) show that Latino Hispanic students received disproportionately few minor infractions and African American students disproportionately many compared to European American students in elementary and middle schools (Skiba et al., 2011). Figure 1 presents the risk ratios for receiving a minor office discipline referral in the classroom for African American and Latino Hispanic students compared to European American students. These risk ratios indicate that, for every 100 European American students who receive a minor office discipline referral, 177 African American and 66 Latino Hispanic students are likely to receive the same.

![Figure 1. Risk ratio of office discipline referral from the elementary classroom.](Image)
This pattern recurs across multiple studies (Wallace Jr, Goodkind, Wallace, & Bachman, 2008). African American students are disproportionately likely to receive minor referrals and major referrals (Bradshaw, Mitchell, O’Brennan, & Leaf, 2010), especially for defiance (Gregory & Weinstein, 2008). African American and Latino Hispanic students are more likely to receive a major office discipline referral for disrespect (Kaufman et al., 2010).

Researchers continue to study the reasons for disproportionality in discipline (Carter, Skiba, Arrendondo, & Pollock, 2017). Theories explaining unequal discipline practices point to the influence of racism (Carter et al., 2014), implicit bias (Pearson, Dovidio, & Gaertner, 2009), cultural discontinuity or mismatch (Gregory & Weinstein, 2008; Tyler et al., 2008; Vavrus & Cole, 2002), stereotypes (Dovidio & Fiske, 2012; Stevenson, 2008), and poverty (Wallace et al., 2008). Though discussion of these various theories exceeds the scope of this review, it is to be noted that the factors contributing to disproportionality in school discipline remain complex and deserve further research.

For the purpose of this study, the critical element of disproportionate school discipline occurs at the vulnerable decision points involving the discretionary responses to subjective interpretations of student behavior (McIntosh et al., 2014; Pearson et al., 2009). The first vulnerable decision point occurs when a teacher decides to address challenging behavior with an office discipline referral and the next as the administrator decides on consequences. In effect, it is at the juncture of these vulnerable decision points that exclusionary practices originate, contributing to disproportionately adverse outcomes for racially and ethnically diverse students.
**Disproportionality in special education.** As it does in school disciplinary practices, disproportionality occurs among students who are deemed eligible for special education services in subjectively-defined categorical disabilities (Anyon, 2009). Socially determined categories of educational disabilities, such as emotional disturbance or a specific learning disability, rely more heavily on professional judgment than do the more biologically informed disabilities such as hearing impairment (Donovan & Cross, 2002). Perhaps as a result, disproportionality occurs more frequently and to a greater degree in categorizing students for these disabilities than for more medical disabilities (Connor & Ferri, 2005; Klingner et al., 2005). The differences in the risk ratios across racial and ethnic groups illustrate the vulnerability of professional judgment to bias (Hosp & Reschly, 2003).

Historically, racially and ethnically diverse students have experienced disproportionality in the selection for special education. As president of the Council for Exceptional Children, Lloyd Dunn (1968) asked whether the disproportionality of racial and ethnic groups in subjectively-determined categories of special education was justifiable. Since 1974, the Office for Civil Rights (OCR) in the U.S. Department of Education has collected data regarding racial and ethnic composition of students participating in special education services. Studies of the OCR data consistently demonstrate over-representation and under-representation (Chin & Hughes, 1987; Donovan & Cross, 2002; Finn, 1982; Oswald, Coutinho, Best, & Singh, 1999). The over- and under-representation of diverse racial and ethnic students in special education programs continue to perplex educators to this day (Civil Rights Data Collection, 2016).
A pattern of disproportionality in special education in the more subjectively defined categories that was identified in the *Digest of Education Statistics 2015* appears in Figure 2 (Snyder, de Brey, & Dillow, 2016). As the reliance on subjective judgment increases in determining who is eligible for special education, so does the range of over- and under-representation of racially and ethnically diverse students. For example, under the more medically defined category of hearing impairment, out of the six racial and ethnic groups, only Pacific Islanders show a disproportionate risk level relative to European American students. Yet, African American, American Indian, and Native Alaskan students continue to appear overrepresented in the most subjectively gauged categories of emotional disturbance, specific learning disability, and intellectual disability. Alternatively, Asian American students appear underrepresented across all subjective disabilities. These patterns of disproportionality persist in studies using contemporary multi-level modeling methods that control for poverty and gender (Sullivan & Bal, 2013; Zhang, Katsiyannis, Ju, & Roberts, 2014).

*Figure 2. Risk ratio for eligibility for special education services.*
Directly pertinent to the current study, the data for consideration for special education eligibility also suggest that increasing the use of objective data may reduce the level of disproportionality. That is, the more special education decisions rely on measurement rather than subjective judgment, the less likely disproportionate decisions will be influenced by a student’s perceived race or ethnicity. Specifically, five of the six racial and ethnic groups exhibit disproportionality under the subjective category of emotional disturbance compared to one of six groups under the medical category of hearing impairment. Whereas the eligibility for emotional disturbance relies heavily on subjective judgment, the eligibility for hearing impairment depends on the measurements of hearing loss instruments. As the weight of objective data increases in the determination of eligibility, the frequency and magnitude of disproportionality decreases.

**Outcomes of disproportionality in school discipline and special education.**

Unfortunately, short- and long-term educational outcomes for students in special education under the category of emotional disturbance (Wagner, Kutash, Duchnowski, & Epstein, 2005) and for those with high levels of disciplinary actions are far from positive (Noltemeyer, Ward, & Mcloughlin, 2015). Because it impacts a student’s sense of belonging at school (Bottani, Bradshaw, & Mendelson, 2016; Okonofu, Walton, & Eberhardt, 2016), exclusionary discipline contributes to lower educational achievement (Arcia, 2006), disengagement from school (Skiba et al., 2011), and increased probability of dropping out (Lee, Cornell, Gregory, & Fan, 2011; Suh, Suh, & Houston, 2007). Additionally, a suspension based on a subjective interpretation of a student’s behavior, such as insubordination, may triple that student’s likelihood of becoming involved in the juvenile justice system (Fabelo et al., 2011). According to American Bar Association,
the phrase “school-to-prison pipeline” originally referred to the sequence associated with excessive school discipline, school drop-out, and subsequent involvement with the judicial system (Redfield & Nance, 2016). Thus, disproportionality in school discipline contributes to the school to prison pipeline (Darensbourg, Perez, & Blake, 2010; Mallet, 2016; Skiba, Arredondo, & Williams, 2014).

The behaviors that lead teachers to make referrals for discipline and for special education may actually reflect unmet mental health needs (Bogart et al., 2013). For example, Harrison, Vannest, Davis, and Reynolds (2012) analyzed elementary teachers’ ratings of 120 behavioral items and found that the most prevalent problem behaviors included excessive worry, distractibility involving low concentration and excessive movement, talking without permission, and not following directions. Similar results were observed in a survey in which teachers rated the top three classes of mental health issues as disruptive behaviors (e.g., getting out of a seat, talking out of turn, arguing, failing to comply with rules), lack of motivation to learn, and disrespectful behaviors (Walter, Gouze, & Lim, 2006). Thus, unrecognized mental health concerns may underlie the problem of disproportionality in school discipline and special education.

Teachers do recognize that students exhibiting challenging behaviors also experience mental health problems. Social emotional learning in school promotes mental health through social relationships and personal coping skills (Durlak, Weissberg, Dymnicki, Taylor, & Schellinger, 2011). However, some students may not achieve developmental milestones and social emotional learning. These students have trouble developing and maintaining personal relationships or responding appropriately to
stressors at school to a degree that interferes with learning. Moreover, these overlooked mental health problems may worsen into mental disorders.

**Internalizing and Externalizing Spectra of Mental Health Problems**

Mental health problems appear along broad dimensions of internalizing and externalizing problems (Brown & Barlow, 2005; Widiger & Samuel, 2005). Within either internalizing or externalizing spectra, the presentation of symptoms from associated disorders may range from subclinical levels to several disorders. These subclinical disorders fall short of the threshold for any specific diagnosis as defined by the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition*, which lists the criteria for clinical levels of mental disorders (DSM-5; American Psychiatric Association, 2013).

Symptoms vary for these two categories of mental health problems. Internalizing mental problems refer to the broad inward expression of anxious, depressive, withdrawn, and somatic complaints (Ask et al., 2016), a spectrum that encompasses the co-occurrence of anxiety and mood disorders (McLaughlin & King, 2015). Externalizing mental health problems refer to the outward exhibition of impulsivity, irritability, aggression, and defiance of social norms (Olson et al., 2013; Walton et al., 2011), with a spectrum that includes combinations of attention deficit/hyperactivity disorder, oppositional defiant disorder, and conduct disorder (Beauchaine & McNulty, 2013; Tackett, 2010). Framing mental health problems within the spectra of internalizing and externalizing problems helps to describe their dimensional nature without the need for categorical diagnosis (Blanco et al., 2015; Carragher et al., 2015; Coghill et al., 2012; Eaton, 2015).
Even when individuals do not fulfill all the criteria of the clinical level, the distress and functional impairment students experience at subclinical levels often compromise their quality of life (Bertha & Balázs, 2013; Hong et al., 2014). Subclinical levels of mental disorders also increase the likelihood of experiencing more than one disorder. In a study on comorbidity, 52.5% of adolescents who met the full criteria for one disorder also exhibited another disorder at the subclinical level (Lewinsohn et al., 2004). Of these adolescents with one full and another subthreshold disorder, 40% suffered a third subclinical disorder. Similarly, 65.8% of youth who met the full criteria for generalized anxiety also experienced subthreshold levels for a second anxiety disorder (Burstein, Beesdo-Baum, He, & Merikangas, 2014). These studies of subthreshold disorders suggest that youth experience mental disorders along a continuum of subclinical to severe along the spectrum of internalizing and externalizing disorders.

**Prevalence of mental health disorders in school-age children.** Cumulative prevalence describes the proportion of persons affected by mental health problems compared to the total sample at any time up to the survey. The results of the National Comorbidity Survey Replication Adolescent Supplement (NCS-A) establish the initial magnitude of the prevalence for mental disorders experienced by youth (Merikangas et al., 2010). According to this study, at some point before age 18, nearly one half of youth will suffer a diagnosable mental disorder. Furthermore, 22% will have experienced significant impairment that interferes with their functioning in school, family, or community. Teenagers suffer from high cumulative rates of anxiety (31.9%), disruptive behavior disorders including ADHD (19.6%), and depressive mood disorders (14.3%). For internalizing disorders among children and youth, the cumulative prevalence
improves the likelihood of accurate identification because many of the cognitive and affective symptoms of adolescents show up initially as somatic complaints when they are children (Ask et al., 2016; Shanahan et al., 2015).

The point prevalence measures the proportion of persons affected by mental health problems compared to the total sample at the time of the survey. In particular, the annual prevalence establishes the percentage of the school-age population that meets the criteria for a mental disorder based on level of functional impairment. For some children who meet the criteria for the disorder with only mild functional impairment, the disorder may nonetheless noticeably affect educational performance. For other children, however, who experience moderate to severe impairment, the adverse effect severely compromises their functioning at home and school (Merikangas, 2011). The annual prevalence rates from the Center for Disease Control and Prevention (CDC) indicate that 13 – 20% of children and youth between ages 5 and 17 experience a mental disorder (Perou et al., 2013). The results of the 2007 National Survey of Children’s Health (NSCH; Child and Adolescent Health Measurement Initiative, n.d.) indicate that the annual prevalence for children between ages 6 and 11 includes 2.9% with anxiety, 1.4% with depression, 3.8% with conduct disorder, and 7.5% with ADHD. However, the NSCH relies on parent reports, and many culturally and linguistically diverse families access no health care for mental disorders, resulting in under-reported diagnoses (Delphin-Rittmon et al., 2013).

Mental health problems in children are often overlooked, with the age of initial diagnosis often depending on the severity of symptoms. Children with more severe presentations are identified sooner than those with milder ones. According to the 2011 NSCH (Visser et al., 2014), parents reported that their children with severe ADHD
received an initial diagnosis around age 5.1, those with moderate ADHD by age 6.9, and those with mild ADHD by age 7.8. The median age of onset was at age 6 for anxiety disorders, 11 for behavior disorders, and 13 for mood disorders (Perou et al., 2013).

The severity of mental health problems that begin during elementary school years increase dramatically over time. For example, among children between ages 6 and 12, anxiety disorders increase from 14 to 27.5%; behavior disorders rise from 8.5 to 15.5%; and mood disorders from 1 to 8% (Perou et al., 2013). Overall, the prevalence of mental disorders increases nearly two-fold during the elementary school years.

In sum, the point prevalence provides a base rate for the expected percentage of students in elementary school who meet the criteria for a mental disorder and demonstrate moderate to severe levels of functional impairment (Forness et al., 2012). The meta-analysis of nine prevalence studies involving children and youth (ages 0 - 17) included five studies in which African American or Latino Hispanic groups comprised at least 49% of the sample (Forness et al., 2012). These five studies, conducted between 1996 and 2007, found a range of point prevalence from 11.1 to 21.1% for severe impairment. These point prevalence rates indicate that, in most public elementary school classrooms comprising 20 to 25 students, teachers may expect to find between two and five students with a mental disorder that exerts a severe impairment on their functioning at home or in school (Forness et al., 2012).

**Educational outcomes of externalizing and internalizing problems.**

Externalizing and internalizing problems may compromise a student’s ability to reach his or her academic potential (Breslau, Miller, Chung, & Schweitzer, 2011; Esch et al., 2014; Mojtabai et al., 2015). Students with externalizing disorders often receive poor grades as
well as higher rates of disciplinary removals from school (Reinke, Herman, Petras, & Ialongo, 2008) and are more likely than unafflicted peers to drop out of high school (Darney, Reinke, Herman, Stormont, & Ialongo, 2013; van der Ende, Verhulst, & Tiemeier, 2016). Students with internalizing disorders exhibit higher absenteeism and lower academic achievement compared to peers without mental disorders, which, in turn, increases the likelihood of drop out (Breslau, Lane, Sampson, & Kessler, 2008; Riglan, Petrides, Federickson, & Rice, 2014; Rumberger & Lim, 2008). Experiencing a mental health problem places a heavy burden on school-aged children’s daily functioning (Erskine et al., 2015), one that snowballs over the course of years as students move through the school system (Masten et al., 2005).

**Disproportionate Unmet Mental Health Needs**

These prevalence rates among elementary students suggest that mental health problems, subclinical to severe, affect every classroom. Manifestations of such problems often result in a referral for school discipline rather than mental health care (Marrast, Himmelstein, & Woolhandler, 2016). Yet, less than 1% of the national population of public school students receives special education services under the category of emotional disturbance (Forness et al., 2012; Kena et al., 2016). The gap between the CDC prevalence and the NCES data on the percentage of students served in special education suggests the under-identification of students needing social, emotional, and behavioral supports (Kaufman, Mock, & Simpson, 2007). Students from racially and ethnically diverse backgrounds receive 70 to 80% of all mental health services from school-based services unrelated to special education (Locke et al., 2017).
Unmet mental health needs may place a disproportionate burden on racially and ethnically diverse students. Studies indicate that only 11% to 54% of children with a mental disorder receive treatment anywhere at all (Jensen et al., 2011; Simon et al., 2015). Moreover, only one in three students with a mental disorder participate in intervention services at school (Simon et al., 2015). This gap in meeting mental health needs places an even greater burden on racially and ethnically diverse students (Gudiño, Lau, Yeh, McCabe, & Hough, 2009). Specifically, nearly one in three Latino Hispanic children needing intervention and approximately one-half of both African American and Asian American/Pacific Islander in need receive mental health services (Kataoka, Zhang, & Wells, 2002; Office of the Surgeon General, 2001). Thus, the problem of disproportionality in school discipline and special education may mask the unmet mental health needs of racially and ethnically diverse students.

**National Recommendations on Universal Screening for Mental Health Problems**

The call for schools to provide universal screening for early identification and prevention of mental health problems has been increasing. Government agencies and policy makers have recommended mental health screening as a means to address unmet mental health needs through early identification in hopes of reducing the burden on children and families. The following is a list of reports from agencies and committees along with legislative actions naming some of the national recommendations for universal mental health screening: (a) the Surgeon General’s Report, *Mental Health: Culture, Race and Ethnicity* (US Department of Health and Human Services, 2002); (b) the President’s Commission on Excellence in Special Education (Branstad et al., 2002); (c) the Committee on Minority Representation in Special Education from the National
Research Council (Donovan & Cross, 2002); (d) the President’s New Freedom Commission on Mental Health (New Freedom Commission, 2003); (e) the re-authorization of the Individuals with Disabilities Education Improvement Act of 2004 (U.S. Department of Education, idea.ed.gov, January, 2007); and (f) the Institute of Medicine’s Committee on the Prevention of Mental Disorders and Substance Abuse among Children, Youth, and Young Adults (O'Connell et al., 2009).

Most recently, the re-authorization of the Every Student Succeeds Act (ESSA) has recommended mental health screening as a component of multi-tiered systems of support in activities intended to support safe and healthy students (Every Student Succeeds Act of 2015, Pub. L. No. 114-95 § Section 4108 (2015-2016)). Multi-tiered systems of support, as a service delivery model, emphasize prevention and early intervention as an alternative to the traditional wait-to-fail model of identification (Brown-Chidsey & Bickford, 2016).

**Traditional teacher referral.** If disproportionality in school discipline and special education is associated with the subjective interpretation of student behavior, then teacher assessments of behavior may need to be better supported. A decision-making aid such as a mental health screening may increase the accuracy of teacher judgment. Further, if disproportionality appears in epidemiological surveys of mental health, then mental screening measures should be tested for evidence indicating whether they contribute to disproportionality through bias.

**Tolerance theory of teachability.** Gerber (1988) postulated that a teacher’s level of tolerance for diversity may contribute to disproportionality in special education. Tolerance theory (Gerber & Semmel, 1984) originates from the behavioral economics model of the teacher as a rational decision-maker who allocates a finite supply of
instructional resources to optimize student learning (Brown & Saks, 1981, 1987). Competing demands from students for a finite array of instructional resources compels the teacher to make decisions regarding the most efficient allocation of resources among students who respond differently to the teacher’s efforts (Gerber, 2005). In an oversimplified illustration, Brown and Saks (1981) compared the learning curves of students A and B. Student A learns more efficiently than student B based on the same amount of time and effort from the teacher. On one hand, if the teacher provides more instructional effort to enable student B to attain proficiency, then student A will not profit as much. On the other hand, if the teacher devotes the same instructional effort to both, student B falls farther behind.

Brown and Saks (1987) extended the two-student model to a more complex model of classroom teaching that presents two potential outcomes. Under the “elitist” option, the teacher may choose to allocate resources to the group of students who show the most promise to profit from the teacher’s instructional effort, resulting in the greater gains in the class mean. Alternately, the teacher may choose the “leveling” option by distributing resources to narrow the range of student performance and reduce the variance. Gerber and Semmel (1985) claim that, when teachers allocate instructional resources, they must choose between maximizing the mean outcomes and minimizing the variance; they cannot choose both.

In an effort to optimize instructional efficiency, teachers tend to orient their finite instructional resources toward the group of students with the most similar educational needs (Gerber & Semmel, 1984), thereby creating a modal group with homogenous academic and behavioral characteristics. The variance of the modal group, compared to
the range of whole class, creates a sense of normality (Gerber, 1988). Tolerance refers to the teacher’s range of accepted academic and behavioral diversity in the context of a classroom. Those students who respond well to instructional resources, fall within the teacher’s tolerance and are thus ascribed the characteristic of “teachability” (Gerber, 1988; Gerber & Semmel, 1984).

As an individual student deviates from the teacher’s sense of normality within the bounds of the modal group, the teacher perceives the student as excessively difficult to teach or manage. At some point, the student surpasses the teacher’s idiosyncratic threshold for tolerance and becomes perceived as intolerable (Gerber & Semmel, 1984), especially students with externalizing problems. In such cases, the teacher often refers the student for either disciplinary action in order to remove him/her from the classroom or to special education in an effort to reduce the strain on the teacher’s instructional effort (Gerber & Semmel, 1985). By contrast, teachers are unlikely to notice students with internalizing problems since their behavior does not infringe on the threshold of tolerance.

Gerber and Semmel (1985) suggested mass (universal) screening to capture students before their behaviors surpass the idiosyncratic threshold of tolerance. The subjective nature of the teacher tolerance may contribute to disproportionality (Gerber, 1988; Shinn et al., 1987). Biases occur to varying degrees, not only for tolerance of specific student behaviors but also for general student characteristics such as gender, racial, or ethnic background (Gerber & Semmel, 1984; Lane, Pierson, & Givner, 2003; Lane, Wehby, Cooley, 2006; Walker, 1985). Biases based on factors unrelated to student
educational needs generate a problem in the subjective boundaries of tolerance for behavioral diversity.

Cultural differences may contribute to over- and under-referring students for assistance. For example, teachers often perceive Asian American children as less sociable compared to other pan-ethnic groups, and so overlook their social and emotional needs (Chang & Demyan, 2007). Furthermore, the social and emotional needs of Asian American students may be obscured because their “model minority” status masks withdrawn behavior (Wing, 2007). Teachers have also referred for special education evaluation based on a non-standard walking style associated with African American culture that departs from dominant cultural norms of strolling (La Vonne, McCray, Webb-Johnson, & Bridgest, 2003). African American students exhibiting higher levels of cultural verve often receive more referrals for discipline as well (Boykin, Tyler, & Miller, 2005). In short, the traditional model of teacher referral relies on the teacher’s individual conception of teachability in addition to a variety of factors unrelated to the mental health needs of each individual student.

The tolerance theory underscores the critical role of the teacher’s professional judgment in the traditional model of identifying students in need of additional support. The strength of their professional judgment comes from the sheer volume of the information that teachers gain from daily interactions with their students. These interactions provide the material for personal observations, which form the basis for inferences, and, ultimately, professional judgments (Groves & Meehl, 1996).

Referrals for discipline and special education comprise the most common tools teachers use for managing students who surpass the subjective threshold of teachability,
as few others are available to them. As the saying goes, “when you only have a hammer, everything looks like a nail.” If universal mental health screening were added to their toolboxes, teachers would gain a valuable method for considering students’ social and emotional needs prior to imposing punitive discipline or categorical tools from special education. Universal mental health screening offers an alternative to the subjective threshold of the traditional model of teachers making referrals.

**Universal screeners enhance professional judgment and decision making.**

Universal mental health screening procedures add structure to a teacher’s professional judgment based on common behavioral criteria (Severson, Walker, Hope-Doolittle, Kratochwill, & Gresham, 2007). This structure allows teachers to better identify students who fall within the range of teachability yet are at risk of suffering mental health problems and eventual disorders. The statistical approach to universal screening—using both risk factors for mental disorders and protective factors for mental health—relies on empirically derived rules for more objective and accurate decision-making. While professional judgment relies solely on the subjective integration of direct observation and interactions, statistical approaches use evidence-based rules for interpreting scores from measurement instruments applied to the student (Garb, 2005).

Statistical approaches improve professional decision-making by providing a more objective probability of identifying subclinical to clinical levels of mental health problems, building upon observable indicators to improve the judgment of laypersons untrained in mental health (Meehl, 1954/1996). Two independent meta-analyses have found that added precision in predicting outcomes resulted when measurement instruments, such as screeners and behavior rating scales, contributed to decision-making
(Ægisdóttir et al., 2006; Grove, Zald, Lebow, Snitz, & Nelson, 2000). In these studies, mental health clinicians (e.g., social workers and psychiatrists) made more accurate decisions regarding diagnosis and prediction based on statistical approaches compared to clinical judgment. Although teachers receive little, if any, training in mental health (Koller & Bertel, 2006; Reinke, Stormont, Herman, Puri, & Goel, 2011), these studies lend support to teachers’ use of empirically based approaches to decision-making in the form of structured professional judgment (Franklin, Kim, Ryan, Kelly, & Montgomery, 2012).

Structured professional judgment improves upon professional judgment through the universal mental health screening procedures and empirically derived thresholds of levels of risk. Teachers’ professional judgment draws from a wide array of observations and interactions with students (Walker et al., 2015). The screening procedure, by contrast, focuses the teacher’s consideration on a few key indicators of risk and resilience that have demonstrated empirical reliability and validity for the identification of current concerns and prediction of probable outcomes. The teacher considers the screening information for any individual student in light of his or her experiential impressions in a filtered manner that improves the accuracy of identifying mental health problems. Ultimately, the teacher still makes the decision regarding the level of risk, while the screening enhances the reliability and validity of that decision (Swets, Dawes, & Monahan, 2000).

**Universal mental health screening procedure.** Universal mental health screening typically begins with gathering information about all the students in a school using a brief assessment, one that concurrently indicates whether students need additional
support based on the prediction of future outcomes (Albers & Kettler, 2014). Next, a teacher—typically a homeroom teacher—rates all the students in their classroom (Ikeda, Neessen, & Witt, 2008; Oakes et al., 2014). The likelihood of universal screening being implemented in schools increases when the instruments that are available are free of charge (Severson et al., 2007) and require low amounts of time and effort (Lane, Oakes, Menzies, 2010). Besides proving feasible in terms of time, money, and effort, screeners must demonstrate reliability, validity, and fairness (Glover & Albers, 2007).

**Psychometric properties of the scores from screeners.** The psychometric properties of reliability and validity help to ensure fair decision-making based on screening instruments (Messick, 1998). Although the validity of a measurement instrument extends beyond its psychometrics (Kane, 2013), the psychometric qualities of the measurement properties influence the inferences drawn from the scores (Zumbo, 2005). The characteristics of the sample of participants selected for the development of the screening instrument largely determine its psychometric properties of reliability and validity. Ideally, the composition of the sample should represent the diversity of the population to be screened (Anastasi & Urbina, 1997; Groves et al., 2009). When the sample does not reflect the diversity of the screened population, the variability in the attenuation for each ethnic group may lead to inaccurate inferences across the racial and ethnic groups. Lower reliability due to measurement error attenuates reliability coefficients and distorts the meaning of the data from which teachers make inferences (Raykov, 2012).

The most common understanding of construct validity refers to the degree to which an instrument actually measures the construct or concept that it purports to
measure (Cronbach & Meehl, 1955). In particular, generalizability and consequential validity play an important role in providing evidence that scores from a screening instrument enhance the fairness of a teacher’s professional judgment (Messick, 1995).

Generalizability refers to the extent to which the scores demonstrate consistent meaning across settings and groups (Messick, 1998). The generalizability of the scores requires empirical evidence of their shared meaning across settings and groups. The current study is concerned with the need for shared meaning of the constructs across racial and ethnic groups. Structural validity refers to whether the construct of the instrument influences the scores on the measure (Borsboom, Mellenbergh, & van Heerden, 2004). According to Messick (1995), “Construct irrelevant variance is always a source of invalidity in the assessment of construct meaning and its action implications.” (p. 753). Scores with poor structural validity lead to erroneous interpretations, and thus lack consequential validity.

Consequential validity refers to the implications of the actions taken as a result of the inferences made from the scores. These implications include the accrual of both incremental positive consequence and the avoidance of unnecessary adverse impact (Messick, 1995). For example, the mental health screeners demonstrate fairness when teachers consistently interpret the scores across students from diverse racial and ethnic groups.

**Validity and fairness.** The *Standards for Educational and Psychological Testing* from the Joint Committee of the American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education (2014) state that fairness means an assessment instrument measures the
intended construct and that the resulting scores are not influenced by factors unrelated to the construct, such as race, gender, or socioeconomic status. In the context of technical adequacy of validity and reliability, a fair test “reflects the same construct for all test-takers, and scores from it have the same meaning for individuals from the intended population; a fair test does not advantage or disadvantage some individuals because of characteristics irrelevant to the intended construct” (AERA, APA, & NCME, 2014, p. 50). Multigroup confirmatory factor analysis provides a test for measurement invariance, meaning the screeners measure the same construct and use the same measurement properties for each represented group. Screeners with measurement equivalence across racial and ethnic groups reduce bias in referrals for assistance (Skiba et al., 2002) and contribute to fair and valid decisions. Ridley, Tracy, Pruitt-Stephens, Wimsatt, and Beard (2008) claim that multicultural assessment validity is the preeminent issue of assessment today.

**Constructs in this Study**

The current study will examine the constructs of teachability, externalizing problems, and internalizing problems. These constructs allow teachers to identify problems as well as strengths. Identifying externalizing and internalizing problems while they are still at a low level offers the opportunity to prevent larger, more difficult problems from developing. Furthermore, identifying students’ level of prosocial skills focuses teachers’ instructional efforts toward those student behaviors they can influence in the classroom. These constructs combine for a more comprehensive approach to assisting teachers in teaching the whole child.
Teachability. Student characteristics that manifest the Gerber and Semmel’s construct of teachability (1984) have been studied recently under the construct of academic enablers (DiPerna & Elliott, 2000). Teachability refers to a class of student behaviors that enable students to participate in and benefit from instruction (DiPerna & Elliott, 2002; Gerber, 1988). These academic enablers include interpersonal skills, motivation, engagement, and study skills (DiPerna & Elliott, 2000). Prior academic achievement and interpersonal skills influence motivation, which in turn influences engagement and study skills (DiPerna, Volpe, & Elliott, 2001).

Interpersonal skills refer to the set of social skills that enable the student to cooperate with others (Wentzel & Watkins, 2002). The competent use of these prosocial skills makes positive social interactions more likely and reduces the effects of negative interactions (Domitrovich, Durlak, Staley, & Weissberg, 2017). Gresham and Elliott (2008) claim that teachers rate the following social skills as necessary for teachability: listening to others, following directions, following classroom rules, ignoring distracting peers, seeking help, taking turns in conversations, cooperating with others, acting kindly with others, acting responsibly, and controlling temper during conflict.

Motivation appears as the willingness to initiate and sustain effort toward reaching a goal (DiPerna, 2006; Linnenbrink & Pintrich, 2002). Students direct their effort toward incremental learning (Cain & Dweck, 1995). Dweck (2002) outlined the developmental changes in motivation based on conceptions of fixed and growth mindsets in students from kindergarten through fifth grade. Initially, kindergarten students strive toward a mastery standard of learning, indicating a growth mindset. Second-grade students develop interest in normative standards of learning, indicating a fixed mindset.
By fifth grade, students adopt a tendency toward either a fixed or growth mindset. These mindsets strongly affect motivation toward engagement in learning. Students demonstrate academic engagement by orienting to the teacher and task at hand, following directions, and participating in discussions (Greenwood, Horton, & Utley, 2002).

Students exhibit study skills with their intentional engagement of cognitive and metacognitive strategies and skills (DiPerna & Elliot, 2000; Gettinger & Seibert, 2002). Study skills involve self-regulated learning strategies that result in deeper information processing (Richardson, Robnolt, & Rhodes, 2010; Schunk & Zimmerman, 1997).

Cognitive skills, such as organizing, synthesizing, and applying knowledge, support efficient learning. Students use meta-cognitive skills, such as self-monitoring, to adjust their approach to learning according to the effectiveness of the strategy (Harvey & Chickie-Wolfe, 2007).

**Externalizing problems.** Externalizing problems include antisocial and disrespectful classes of behavior (Gresham, 2015). Antisocial behaviors, such as overt fighting in addition to covert lying and stealing, violate social mores (Achenbach & Rescorla, 2007). Disrespectful behaviors include negative social interactions as students engage in excessive levels of defiance of teachers (Henricsson & Rydell, 2004; Murray & Murray, 2004) and exhibit hostility toward peers (Parker, Rubin, Erath, Wojslawowicz, & Buskirk, 2006).

**Internalizing problems.** Internalizing problems are manifest within the self (Achenbach & Rescorla, 2007). Anxiety and depression share the common feature of covert emotional distress (Chorpita, 2002). However, observable signs of internalizing problems can be detected as frequent somatic complaints and withdrawal.
Students with anxiety problems exhibit emotional dysregulation of anxiety (Weems, 2008). They may display chronic worry through behavioral avoidance or physiological symptoms such as muscle tension (bracing), tremors, increased heart rate, or rapid breathing. Elementary school children with anxiety exhibit specific phobias and separation anxiety between ages five to nine, developing fears of death, danger, failure and social situations between the ages 10 through 13 (Beesdo, Knappe, & Pine, 2009; Weems & Silverman, 2016).

Children experiencing depression often display a low mood, lack of energy, and lack of interest in the activities at school, appearing unmotivated or not engaged in learning or interacting with others (Keenan et al., 2008; Weiss & Garber, 2003). Such students demonstrate a low tolerance for frustration and often cry. Socially, students with depressive symptoms may fluctuate between angry irritability and withdrawal into sadness.

Students with internalizing symptoms often report somatic complaints such as headaches, stomachaches, or muscle pain (Garralda, 2011; Shelby et al., 2013). During the transition to elementary school, the frequency of somatic complaints increases temporarily. Although these somatic complaints lack any medical explanation, students frequently miss school because of them (Janssens, Oldehinkel, Dijkstra, Veenstra, & Rosmalen, 2011).

Student withdrawal appears as spending significant time alone (Rubin, Coplan, & Bowker, 2009). Socially withdrawn students avoid interacting with peers while in their company. Moreover, such students appear to actively shy away from initiating and maintaining social interactions.
Strengths-based mental health assessments complement screening for internalizing and externalizing problems (Suldo & Shaffer, 2008). Taking into account both student risks and strengths improves the intervention utility of the screener (Hayes, Nelson, & Jarrett, 1987; Hunsley, & Meyer, 2003). The assessment of teachability through identifying student strengths in academic enablers permits the teacher to distinguish between skill deficits and performance deficits (Jenkins & DeMaray, 2015). For example, the student who lacks the skills to solve interpersonal problems requires a different intervention than the student who has the skills but lacks the motivation to use them. Strengths-based screeners rate students on the presence of social and emotional skills that enable them to engage in learning (Keyes, 2009; Marquez et al., 2014). Students with higher social and emotional skills develop healthy relationships and higher academic achievement (Davis, Solberg, de Baca, & Gore, 2014; Garner, Mahatmya, Brown, & Vesely, 2014). Specifically, prosocial skills that enable students to benefit from instruction improve the teacher’s perception of teachability (Elliott, DiPerna, Mroch, & Lang, 2004). These constructs of prosocial skills involved in teachability, as well as internalizing and externalizing problems, provide the teacher with a sophisticated lens through which to identify students in need of additional support.

**Mental Health Screeners in this Study**

The screeners in this study address both mental health problems and prosocial skills because the absence of problems does not necessarily imply mental health (Suldo & Shaffer, 2008). The *Elementary Social Behavior Assessment* (ESBA) measures the latent variable of academic enablers. The *Student Risk Screening Scale* (SRSS) measures the
latent variable of externalizing problems. The Student Internalizing Behavior Screener (SIBS) measures the latent variable of internalizing problems.

**Elementary Social Behavior Assessment.** The *Elementary Social Behavior Assessment* (ESBA) presents 12 positively stated items in a strengths-based approach to screening for social skills that contribute to academic achievement (Pennefather & Smolkowshi, 2015). The ESBA measures the construct of teachability—in other words, the skills that (a) strengthen the student-teacher relationship, (b) build friendships, and (c) promote school adjustment and academic success (Gerber & Semmel, 1984; Walker et al., 2015). The original rating system used a 3-point scale in which a 3 indicated *mastery*, 2 suggested the *need for improvement*, and 1 pointed to an *area of concern*. However, in this study, the scaling system has been adjusted from mastery to frequency of skill. In this study the ESBA ratings are 0 for *never*, 1 for *occasionally*, 2 for *sometimes*, and 3 for *frequently*. The sum of the rating on the 12 items comprises the total score, with a range of 0 to 36 points. Scores above 18 indicate low risk and scores below 12 indicate high need for additional support. Pennefather and Smolkowski (2015) conducted a pilot study followed by a validation study of the ESBA. The pilot study included 187 students in kindergarten through grade three in a rural town in the Pacific Northwest. The racial and ethnic composition of the sample consisted of 76.6% European American, 9.5% Latino Hispanic, 4.9% Asian American or Pacific Islander, 1.6% African American, 1.1% American Indian, and 6.3% multiracial students. All of the items correlated with the total score (*r* = .67 to .88) and Cronbach’s alpha for internal consistency was α = .95. The fit indexes from the exploratory factor analysis with ordered categorical variables reached .989 on both the comparative fit index (CFI) and Tucker-Lewis index (TLI). Moreover, a
single factor supported 11 of the items with loadings of .84 and one item with a loading of .76. The ESBA demonstrated concurrent validity \((r = .84)\) with the *Walker–McConnell Scale of Social Competence and School Adjustment – Elementary Version* (WMS; Walker & McConnell, 1995). Furthermore, the ESBA correlated strongly with the teacher-preferred subscale \((r = .84, p < .001)\), the school adjustment subscale \((r = .83, p < .001)\), and the peer-preferred subscale \((r = .73, p < .001)\) on the WMS.

The validation study of the ESBA occurred within a larger study of a classroom management program and did not include demographic data on the students \((N = 1616)\) from California, Oregon, and Washington State (Pennefather & Smolkowski, 2015). With the larger sample, the internal consistency indicated high reliability with \(\alpha = .95\) at pretest and \(\alpha = .94\) at post-test. The test-retest reliability \((r = .76, p < .001)\) remained stable between eight weeks. The confirmatory factor analysis produced a CFI of .98 and a TLI of .97 for a single factor model of teachability. Moreover, the factor loadings spanned a range from .81 to .94. The ESBA has demonstrated adequate reliability and validity for identifying students who exhibit prosocial skills that enable them to participate in and benefit from instruction (Pennefather & Smolkowski, 2015; Walker et al., 2015).

**Student Risk Screening Scale.** Drummond (1994) developed the *Student Risk Screening Scale* (SRSS) to identify students with antisocial behavioral problems. The teachers rate each student on seven items from 0 to 3 according the frequency of observed problem behaviors. The sum of the scores (range 0 – 21 points) indicates the risk status from low (0-3), moderate (4-8), and high (9-21). Drummond, Eddy, Reid, and Bank (1994) established initial predictive validity showing that the SRSS predicted
academic and behavior problems between 1.5 and 10 years later and correlated with the Aggressive Behavior subscale of the Child Behavior Checklist – Teacher Report Form (Achenbach, 1991). More recently, Lane and colleagues have updated studies of reliability and validity for the SRSS (Ennis, Lane, & Oakes, 2012; Lane, Little, et al., 2009; Lane, Kalberg, Lambert, Crnobori, & Bruhn, 2010; Lane, Menzies, et al., 2012; Menzies & Lane, 2010; Oakes et al., 2010). The SRSS demonstrates internal consistency for elementary students, with values of Cronbach’s alpha between $\alpha = .78$ and $.87$ across six studies involving rural, suburban, and urban elementary schools (Ennis et al., 2012; Lane, Kalberg, et al., 2010; Lane, Little, et al., 2009; Lane, Menzies, et al., 2012; Lane, Richards-Tutor, Oakes, & Connor, 2013; Menzies & Lane, 2012). The test-retest reliability for the SRSS between fall and winter is $r = .73$ to $.79$; between winter and spring $r = .74$ to $.86$; and between fall and spring $r = .68$ to $.77$.

Multiple research studies show that the SRSS provides strong support for validity claims. The SRSS demonstrates convergent validity with the Strengths and Difficulties Questionnaire (Goodman, 1997) with $r = .71$ to $.80$ ($p < .0001$) (Ennis et al., 2012) and discriminant validity from prosocial and academic skills on the Social Skills Improvement System- Performance Screening Guide (Elliott & Gresham, 2008) with $r = -0.50$ to $-0.63$ ($p < .0001$) (Lane, Richards-Tutor, et al., 2013). Furthermore, the SRSS negatively correlates with academic achievement on state proficiency exams between $r = -0.46$ and $-0.61$ ($p < .0001$) (Ennis et al., 2012). Whereas Menzies and Lane (2012) found correlation with office discipline referrals ($r = .48$, $p < .0001$), Ennis et al. (2012) found moderate correlations with out-of-school suspensions ($r = .26 - .40$, $p < .0001$). Lane, Richards-Tutor, Oakes, and Connor (2013) surveyed teachers regarding the social
validity of the SRSS, resulting in moderately high 22.14 (\(SD = 7.96\)) out of 32 points possible.

Across five studies in various settings with homogenous and diverse populations, the SRSS demonstrates excellent diagnostic accuracy, with 11 of 15 measurements of the area under the curve (AUC) at, or above, 0.95, for detecting externalizing disorders on the SSBD. That is, the SRSS predicts externalizing problems 45% better than chance. On 6 of 11 measurements, the SRSS also predicted internalizing problems with an area under the curve of .75 or better. In other words, although the SRSS is designed to identify students with externalizing problems, the SRSS also identifies students with internalizing problems 25% better than chance. The SRSS accurately distinguished between high and low risk based on the Systematic Screener of Behavior Disorders (SSBD; Walker & Severson, 1992) with 94% sensitivity and 95% specificity for externalizing problems among students in kindergarten through grade two (Lane, Little, et al., 2009). Furthermore, the SRSS also demonstrated 44% sensitivity and 95% specificity for internalizing problems. The SRSS demonstrates excellent psychometric properties for the purpose of identifying students at risk for developing antisocial patterns of behavior associated externalizing problems.

**Student Internalizing Behavior Screener.** The Student Internalizing Behavior Screener (SIBS) measures the level of observed signs of distress from internalizing symptoms (Cook et al., 2011). Teachers rate every student on seven items using a four-point Likert scale with 0 for *never*, 1 for *occasionally*, 2 for *sometimes*, and 3 for *frequently*. A total score above 8 indicates high risk and below 3 low risk status for developing internalizing problems. Cook and colleagues (2011) established the
reliability and validity of the SIBS in grades one through five in four elementary schools in the Pacific Northwest. The racial and ethnic composition included 48% European American, 20% Latino Hispanic, 13% African American, 6% Asian, 2% Native American, and 11% other. Cronbach’s alpha for internal consistency reached $\alpha = .81$ in the fall and $\alpha = .79$ in the winter. All items correlated with the total score with $r > .50$ (range $r = .51$ to .79). The test-retest correlation coefficient between fall and winter administrations was $r = .74$.

The SIBS demonstrates adequate validity. The SIBS demonstrated convergent validity ($r = .82$) with the internalizing scale of the Child Behavior Checklist - Teacher Report Form (Achenbach, 1991). Moreover, the SIBS correlated with the SRSS with $r = .41$. The SIBS demonstrates strong predictive validity using a cut-score of 8 with area under the curve (AUC) of 0.934. Moreover, the SIBS had an 86% true positive rate with a 1% false positive rate. However, the true negative rate was 99% and the false negative rate was 14%. The SIBS demonstrates adequate reliability and validity toward identifying students with internalizing problems. All of the items on the SIBS correlated with the total score with $r > .51$ (Cook et al., 2011). Whereas sad or unhappy ($r = .79$) demonstrated the strongest relationship with the total score, bullied by peers showed the weakest correlation with the total score ($r = .51$).

Each screening instrument demonstrates adequate psychometric properties. The screeners exhibit high reliability based on internal consistency with Cronbach’s alpha and test-retest stability of scores between successive administrations. Each screener also demonstrates convergent validity with other measures of related constructs and predictive validity toward the accurately distinguishing students who need additional support from
those who do not at the time of administration. In particular, each screener gives the level of sensitivity for correct identification and the specificity for incorrect identification. Each screener demonstrates social validity namely through feasibility of brevity of time costs and availability for zero financial cost. The racial and ethnic composition of the studies’ samples supports cultural validity. The current study builds upon these psychometric properties by testing for measurement invariance across African American, Asian American, Latino Hispanic, multiracial groups of students compared to reference group of European American students.

**Multigroup Confirmatory Factor Analysis**

Multigroup confirmatory factor analysis provides a structural equation modeling approach to address questions regarding structural validity and generalizability of scores across diverse racial and ethnic groups (Hui & Triandis, 1985). It provides a measurement model that identifies the extent of construct variance both relevant and irrelevant (van de Vijver, 2011b). Multigroup confirmatory factor analysis tests the degree of measurement invariance across racial and ethnic groups (Millsap, 2012). If a screener fails to demonstrate measurement invariance, the resulting measurement bias makes interpreting the score difficult. Biased scores may reflect an artifact of measurement (Meredith & Teresi, 2006) that allows students at risk to fall through the cracks (Chen, 2008). Multigroup confirmatory factor analysis takes a step toward the recommendations from the *Standards for Educational and Psychological Testing* to evaluate factor structures of assessments for differential test functioning (AERA, APA, & NCME, 2014).
Measurement Invariance

Measurement invariance refers to the “degree of congruence” between the scores from the instruments and the measured constructs across multiple groups (Haynes, Smith, & Hunsley, 2011). For example, mental health screeners often use an ordinal scale for relative frequency of the occurrence of observed behaviors, such as never, occasionally, sometimes, or frequently. Testing for metric invariance provides evidence for the extent to which each of the racial and ethnic groups shares the same unit of measurement. A common unit of measurement is necessary, but not sufficient, for making comparisons among groups. Thus, establishing a common origin of measurement addresses whether never shares a common starting point for the scale across racial and ethnic groups. Scores that share both a common metric and common scale allow interpretations across racial and ethnic groups.

Configural invariance. The constellation of the indicators that load onto the latent construct creates a pattern of relationships. This pattern specifies causal relationship between the construct and the covariation among the items (Borsboom, 2008). In multigroup confirmatory factor analysis, the pattern of factor loading for the constellation of indicators shares the same factor structure across each group. The pattern of invariance across the comparison groups appears when the each indicator demonstrates a similar strength in factor loading. Furthermore, none of the groups presents a measurement model having an indicator with a trivial (close to zero) loading. Configural invariance across racial and ethnic groups provides evidence of structural validity and a basis for generalizability. Differences in the factor structure for a group suggest that the latent variable does not share equivalent meaning across the groups. If the pattern of
factor loadings from one group differs significantly from any of the comparison groups, then the configural structure for the discrepant group demonstrates construct irrelevant variance that diminishes the validity of that construct across racial and ethnic groups. By establishing structural validity, the multigroup confirmatory factor analysis may proceed to test the measurement properties of the scores. Evidence of configural invariance enables the next stage of analysis of measurement invariance.

**Metric invariance.** Metric variance refers to variance around the precision of the unit of measurement. When the unit of measurement is consistent across groups, the scores demonstrate metric invariance, a property which allows the scores to measure the latent construct. Conversely, if the unit of measurement differs for any setting or group, then the scores lack a consistent unit. If the measured level differs from the true level according to the group membership, this lack of metric invariance leads to inaccurate inferences about the meaning of the unit of measurement, adding measurement bias into the inferences made from the scores.

**Scalar invariance.** Scalar variance refers to variance around the origin point of measurement. When the initial starting point of measurement is consistent across settings and groups, the scores demonstrate scalar invariance. In comparison, scalar invariance with continuous variables tests whether the intercepts are the same across groups thereby allowing factor means to be compared across groups. However, if the intercepts differ, then determining whether differences arise from true differences across groups or from measurement artifacts proves difficult. By analogy, scalar invariance with ordinal data relies on the thresholds between categories. When the thresholds between never, occasionally, sometimes, and frequently are similar across groups, we can
say that differences from the scale score arise from “true” differences in the latent construct.

Establishing measurement invariance involves testing whether all groups share the same unit of measurement and the same scale of measurement. Such measurement invariance in scores provides the evidence of generalizability across racial and ethnic groups. The generalizability of the scores enables teachers to make fair decisions resulting in consequential validity.

**Summary**

According to the annual prevalence data from the Centers for Disease Control, nearly one in five elementary students meets the criteria for a mental disorder (Perou et al., 2013). Moreover, the prevalence of students suffering from subclinical levels of mental problems appears even higher (Simon et al., 2015). Despite the prevalence of mental problems, the mental health needs of many students go unmet (Kataoka et al., 2002; Merikangas et al., 2010). Universal mental health screening, according to researchers and policy makers, offers an effective and time-efficient means to identify those students who would most benefit from social, emotional, and behavioral supports (Every Student Succeeds Act of 2015, Pub. L. No. 114-95 § Section 4108 (2015-2016); Kettler et al., 2014).

Universal mental health screening instruments can detect subclinical mental health problems before they reach the severity of a diagnosable mental disorder, thus impairing students’ ability to function academically and socially (Stiffler & Dever, 2015). The psychometric properties of the screening instruments enhance the ability of teachers to correctly identify students in need of additional support. The greater reliability of the
instrument supports teachers in making more valid referral decisions. By contrast, lower reliability introduces measurement bias into making decisions, which, in turn, adversely affects the referral. For example, measurement bias that relates to the race or ethnicity of the student attenuates the screener’s ability to detect subclinical mental health problems, compromising valid decisions by either over- or under-identifying the level of risk for future outcomes. Conversely, universal mental health screeners with adequate reliability and validity coefficients complement professional judgment for accurately identifying students in need of social, emotional, and behavioral support (Albers & Kettler, 2014; Cook et al., 2010).

This study addresses a gap in the research on the development of universal screeners that demonstrate measurement invariance based on a sample of African American, Asian American, Latino Hispanic, European American, and multiracial students in elementary schools. The three screening instruments include the Elementary Social Behavior Assessment (Pennefather & Smolkowski, 2015), the Student Risk Screening Scale (Drummond, 1994), and the Student Internalizing Behavior Screener (Cook et al., 2011). The first level of research questions for each screener tests whether the screeners measure the same latent construct of interest with the equivalent factor structure consisting of nontrivial indicators. The second level of research questions tests whether the ordinal descriptors of never, occasionally, sometimes, and frequently demonstrate measurement equivalence in the metric unit of frequency on for latent construct of interest for the screener. The third level of research questions addresses whether the descriptor, never, establishes measurement equivalence on the initial threshold of the scale of the construct of interest.
Chapter Three: Methods

The purpose of this study is to test for measurement invariance in the latent variables from three universal mental health screeners using a sample with adequate representation from African American, Asian American or Pacific Islander, Latino Hispanic, European American, and multiracial backgrounds among students drawn from elementary schools. The first of these three screening instruments, the *Elementary Social Behavior Assessment* (ESBA), measures the latent variable of teachability. The second, the *Student Risk Screening Scale* (SRSS), measures the latent variable of externalizing problems. The third, the *Student Internalizing Behavior Screener* (SIBS) measures the latent variable of internalizing problems. The study of measurement invariance examines the equivalence in the meaning of the constructs measured by the screeners and the equivalence in the measurement properties of the screeners across the five groups. If the screeners demonstrate measurement invariance across the five groups, then decisions based on the obtained scores will be more fair and valid (AERA et al., 2014).

Research Design

This study used extant data in an ex post facto cross-group comparative research design to test for measurement invariance in the latent variables on the screening instruments (Knight, Roosa, & Umaña-Taylor, 2009). This study used a descriptive measurement approach to study the accurate measurement of the underlying latent variables of teachability, externalizing problems, and internalizing problems (Wilson & Gochyyev, 2006). The cross-group comparative research design examined the structure and the levels of latent constructs among groups of African American, Asian American or
Pacific Islander, Latino Hispanic, and multiracial students compared to the European American index group (van de Vijver, 2011a).

Participants and Setting

Data describing the school district from which this study’s extant data were drawn situate the context for the selection process for the elementary schools. The participants comprise all of the students in schools that meet the criteria for selection for universal screening (see Appendix A). The extant data about these students met the minimum required sample size.

Sampling method. This study relied on data collected previously by the school district for an ongoing district initiative on social-emotional interventions. The district initiative developed a multi-stage sampling method that specified criteria for school-wide readiness to implement universal mental health screening in every classroom. The specification of the selection criteria for schools and of the demographics of the students clustered within the schools allows the results of this study to generalize to similar populations. The clusters of student data do not meet the assumption of independence because the teacher completes the ratings for every student in the class. Nonetheless, studies of measurement invariance have demonstrated robustness with regard to dependent data (Perdomo, Jones-Farmer, Edwards, & Svyantek, 2014; Rutkowski & Svetina, 2017).

Selection of participating schools. In the initial stage, the primary sampling unit selected schools from the frame of all elementary schools that met the following three criteria: (a) sustained implementation of the social, emotional, and behavioral (SEB) supports with adequate fidelity, (b) a school-wide continuum of SEB supports, and
(c) a school-wide action plan for implementing targeted, small group interventions for students with similar needs. Figure 9 in Appendix A illustrates the selection process for the schools demonstrating readiness to administer universal mental health screening procedures.

The first criterion for school selection involved the sustained implementation of social, emotional, and behavioral supports linked to the districtwide implementation of School-wide Positive Behavior Interventions and Supports (SWPBIS; Horner, Sugai, & Anderson, 2010; OSEP Technical Assistance Center on Positive Behavioral Interventions and Supports, 2015). Further, the social, emotional, and behavioral supports had to be implemented with fidelity. The district uses the School-wide Evaluation Tool (SET; Horner et al., 2004; Sugai, Lewis-Palmer, Todd, & Horner, 2001; Vincent, Spaulding, & Tobin, 2010) to measure basic fidelity of implementation. The SET established the minimum criteria for implementation fidelity with an overall score of 80% on the seven components of SWPBS in addition to 80% of the five elements of teaching expectations. To be considered for inclusion, schools had to have reached the 80-80 rule on the SET for two consecutive years.

The second criterion for inclusion in this study involved the School-wide PBIS Tiered Fidelity Inventory (TFI; Algozzine et al., 2014; McIntosh et al., 2016). The TFI indicates readiness to implement interventions to address the needs of students identified by the social, emotional, and behavioral screeners. Whereas the SET addresses only core features of tier I, the TFI uses the same format, scale, and language to measure the level of implementation of core features of SWPBS across universal tier I, targeted tier II, and intensive tier III interventions and supports. On the TFI, obtaining 80% of the
components for any tier of intervention indicates the minimum level of fidelity for the specific tier. For example, a school could score 90% in tier I, 20% in tier II, and 80% in tier III, resulting in fidelity at tiers I and III. The TFI provides a more complete picture across the continuum of interventions than the SET. The criterion for inclusion was a complete TFI, though no specific score was necessary.

The third criterion for inclusion in this study was the use of a school-wide action plan for addressing mental health needs. Schools are required to have an action plan for implementing a range of targeted tier II interventions for students identified as needing additional social, emotional, and behavioral support. The tier II plan may include interventions such as inclusion in social skills groups, lunch or reading buddies, or checking in with a caring adult. A range of social, emotional, and behavioral supports should be available to meet the students’ needs if the screening data provides intervention utility (Hayes et al., 1987). The action plan describes the interventions to be implemented and the data to inform the selection of students for supports.

**Selected schools.** Denoted with pseudonyms of various countries, ten schools from 36 total elementary schools met the criteria for inclusion. Table 1 presents the selected schools’ scores on the three criteria with country pseudonyms corresponding to schools. Though every school easily satisfied the criteria for the SET, only three schools provided evidence of multiple tiers of intervention on the TFI. These data suggest that most schools have established only the basic implementation of school-wide positive behavior interventions and support.
Table 1

\textit{Inclusion Criteria for Selected Schools}

<table>
<thead>
<tr>
<th>Schools</th>
<th>Schoolwide Evaluation Tool</th>
<th>School-wide PBIS Tiered Fidelity Inventory</th>
<th>School-wide Action Plan</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Percent of Implementation of Schoolwide Elements</td>
<td>Percent of Key Elements Implemented in</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Expectations Taught</td>
<td>80-80 Rule Met</td>
<td>Tier I</td>
</tr>
<tr>
<td>Belgium</td>
<td>100</td>
<td>100</td>
<td>Yes</td>
</tr>
<tr>
<td>Belize</td>
<td>93</td>
<td>90</td>
<td>Yes</td>
</tr>
<tr>
<td>Denmark</td>
<td>91</td>
<td>90</td>
<td>Yes</td>
</tr>
<tr>
<td>Ecuador</td>
<td>97</td>
<td>100</td>
<td>Yes</td>
</tr>
<tr>
<td>France</td>
<td>90</td>
<td>80</td>
<td>Yes</td>
</tr>
<tr>
<td>Laos</td>
<td>95</td>
<td>100</td>
<td>Yes</td>
</tr>
<tr>
<td>Liberia</td>
<td>94</td>
<td>90</td>
<td>Yes</td>
</tr>
<tr>
<td>Maldives</td>
<td>94</td>
<td>100</td>
<td>Yes</td>
</tr>
<tr>
<td>Poland</td>
<td>95</td>
<td>100</td>
<td>Yes</td>
</tr>
<tr>
<td>Poetry</td>
<td>97</td>
<td>100</td>
<td>Yes</td>
</tr>
</tbody>
</table>
The selected schools comprise higher rates of racial and ethnic diversity compared to the national population of elementary students. Enrollment in selected schools during the 2015-16 academic year included 31.5% European American students and 68.5% students from diverse racial and ethnic backgrounds.

Table 2
Selected Schools and National Elementary School Racial and Ethnic Demographics

<table>
<thead>
<tr>
<th>Race or Ethnicity</th>
<th>Percent</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Schools</td>
<td>National</td>
<td></td>
</tr>
<tr>
<td>American Indian or Alaskan Native</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Asian American or Pacific Islander</td>
<td>13.6</td>
<td>5.3</td>
</tr>
<tr>
<td>African American</td>
<td>18.4</td>
<td>15.6</td>
</tr>
<tr>
<td>Hispanic or Latino</td>
<td>25.4</td>
<td>25.4</td>
</tr>
<tr>
<td>Multi-Racial</td>
<td>11.1</td>
<td>2.8</td>
</tr>
</tbody>
</table>

The school district presents high likelihood of selecting a sample of diverse students from schools that will resemble the national population. The sample frame includes all of the elementary schools within a school district. The Asian and Hawaiian-Pacific Islander groups combine for 11.6% of the total elementary enrollment, while multiracial students identify with two or more races. This study excludes American Indian and Native Alaskan students because the sample size was too small for statistical analysis.

Few of the selected schools resemble the national percentages of students participating in programs for English language learners, special education, and subsidized meals. Only “Poland” appears within 10% of the national percentage of English Language Learners. Ecuador, Liberia, and Maldives approximate the national percentage
of students participating in special education. The percentage of students attending Denmark and Poland resemble national levels of students receiving free or reduced priced meals.

*Figure 3.* School sample and national population participation in selected programs.

Students missed classroom instruction due to exclusionary discipline or not completing grade in the school where they started the year (Figure 4). The percentage of third grade students in the Maldives and fifth graders in Liberia approximate the national level of students suspended or expelled within the 2015-2016 school year. In third grade, France, Laos, and Liberia exceed the national level. In fifth grade, only Poland and Denmark did not suspend any students. The percentage of students suspended in France or Laos approaches between two-to-five times the national level. The average percentage
of all school except France or Laos is lower than national level. Denmark, Ecuador, and France approximate levels of annual student mobility. Belgium, Laos, and the Maldives surpass the national average of student mobility.

**Figure 4.** School sample and national population interrupted participation in instruction.

Students missed classroom instruction due to not completing the grade in the school where they started the year. Denmark, Ecuador, and France approximate average levels of annual student mobility. Belgium, Laos, and the Maldives surpass the national average of student mobility.

**Demographics of the teachers and administrators.** Just over half of the administrators are women. The administrators include eight European Americans, three African Americans, and one Asian American. Just over half of the administrators have
worked in leadership for five years or less and also with SWPBS for one year or less.

Thus, implementing SWPBS is new for most of the leadership. From the complete group of teachers, only ten are male and twenty-one are from racially and ethnic diverse backgrounds. One quarter of the teachers have five years or less teaching experience. Just over half of the teachers have one year or less experience with SWPBS.

Table 3
*Demographics and Educational Experience at the Selected Schools by Percentage*

<table>
<thead>
<tr>
<th>Percent of</th>
<th>Gender</th>
<th>School Administrators (N = 12)</th>
<th>Teachers (N = 161)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>58</td>
<td>94</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>42</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>American Indian</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>8</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>African American</td>
<td>25</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Latino Hispanic</td>
<td>0</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>European American</td>
<td>67</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>Years of Experience</td>
<td>Administrating</td>
<td>Teaching</td>
<td></td>
</tr>
<tr>
<td>0 – 5</td>
<td>58</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>6 – 10</td>
<td>17</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>11 – 20</td>
<td>8</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>21+</td>
<td>17</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>Personal Experience with Schoolwide PBS in Years</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 1</td>
<td>25</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>33</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>2+</td>
<td>42</td>
<td>43</td>
<td></td>
</tr>
</tbody>
</table>

*Selection of student sample.* After schools were selected, the secondary sampling unit included clusters of all of the students attending general education classrooms. Teachers in general education classrooms considered every student on their
class roster for universal screening procedures. Lastly, the total sample from all of the selected elementary schools was stratified into groups based on their racial or ethnic identification according to the school district’s records.

The demographic data for the students screened included the percent of the sample based on gender, grade level, and racial and ethnic group (Table 4). The sample of students comprise nearly balanced between boys and girls and among the six grade levels. The racial and ethnic composition includes higher levels of sampling of Asian American or Pacific Islanders andmultiracial students compared to national percentages.

The African American and Latino Hispanic groups approximate the national enrollment. The European American students appear much less represented in this sample compared to the national population.

Table 4
Demographics of Students by Percentage

<table>
<thead>
<tr>
<th>Gender</th>
<th>N</th>
<th>Sample</th>
<th>National</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>2039</td>
<td>51.4</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>1926</td>
<td>48.6</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Grade level</th>
<th>N</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kinder</td>
<td>661</td>
<td>16.7</td>
</tr>
<tr>
<td>First</td>
<td>661</td>
<td>16.7</td>
</tr>
<tr>
<td>Second</td>
<td>641</td>
<td>16.2</td>
</tr>
<tr>
<td>Third</td>
<td>697</td>
<td>17.6</td>
</tr>
<tr>
<td>Fourth</td>
<td>677</td>
<td>17.0</td>
</tr>
<tr>
<td>Fifth</td>
<td>628</td>
<td>15.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Race and Ethnicity</th>
<th>N</th>
<th>Sample</th>
<th>National</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asian American/Pacific Islander</td>
<td>539</td>
<td>13.6</td>
<td>5.3</td>
</tr>
<tr>
<td>African American</td>
<td>728</td>
<td>18.4</td>
<td>15.6</td>
</tr>
<tr>
<td>Latino Hispanic</td>
<td>1009</td>
<td>25.4</td>
<td>26.4</td>
</tr>
<tr>
<td>Multiracial</td>
<td>441</td>
<td>11.1</td>
<td>2.9</td>
</tr>
<tr>
<td>European American</td>
<td>1248</td>
<td>31.5</td>
<td>48.6</td>
</tr>
</tbody>
</table>

*Note:* Digest of Education Statistics, National Center of Education Statistics, 2016, Table 203.50.
Sample size. A common approach sample size for confirmatory factor analysis involves the Root Mean Square Error of Approximation (RMSEA). The RMSEA method compares the null hypothesis ($H_0$) model and the alternative hypothesis ($H_1$) model based on the degrees of freedom for each model. The degrees of freedom set the critical value of the chi-square distribution for a given alpha significance level for each model (MacCallum, Browne, & Sugawara, 1996). The RMSEA require the degrees of freedom to estimate the sample size. As the degrees of freedom increase, the necessary sample size to establish enough power decreases. Therefore, the model with fewer degrees of freedom (70) was used to calculate the minimal sample size. MacCallum, Browne, and Cai (2006) designed a method for determining the minimum sample needed to test differences in the overall fit between nested models in studies of measurement invariance. The online calculator for the close fit produced each of the sample sizes using the MacCallum RMSEA procedure (Gnambs, 2013). The SRSS and SIBS require a minimum of 838 participants in the sample with power set for 0.80 and an alpha set for 0.05 while adjusting for five comparison groups. When the power is increased to .99, the necessary sample size becomes 1350. The ESBA requires at least 92 in the sample. The extant data in this study exceed the recommended sample size based on the MacCallum RMSEA procedure.

Instrumentation

The screening instruments in this study measure the focal latent variables of teachability, externalizing problems, and internalizing problems. A brief description of the latent construct follows the review of the reliability and validity of each instrument.
The results from previous studies on each of the screeners demonstrate adequate psychometric properties.

**Elementary Social Behavior Assessment.** The *Elementary Social Behavior Assessment* (ESBA) measures the construct of teachability (Gerber & Semmel, 1984; Walker et al., 2015) based on the teacher ratings of the frequency of observed behaviors (Pennefather & Smolkowski, 2015). The ESBA exhibits high internal structure reliability with Cronbach’s alpha in the pilot study (α = .95) and in the validation study during the pre-test (α = .95) and at the post-test (α = .94). Furthermore, the test-retest reliability demonstrated stability in scores over eight weeks (r = .76). The ESBA shows convergent validity (r = .84) with the *Walker–McConnell Scale of Social Competence and School Adjustment – Elementary Version* (WMS; Walker & McConnell, 1995).

Teachability is a class of academic enablers that establishes the basis for learning from classroom instruction (Elliot, DiPerna, Mroch, & Lang, 2004). DiPerna and Elliott (2000) identified four specific academic enablers: interpersonal skills, study skills, motivation, and engagement. Interpersonal skills facilitate cooperative learning and interacting with others (DiPerna & Elliot, 2000). Interpersonal skills include getting along with others, resolving peer conflicts, having normal conversations without becoming hostile, and working out strong feelings appropriately. Additionally, a student demonstrates interpersonal skills when she or he follows rules, avoids breaking rules even when encouraged by peers, and behaves appropriately outside of the classroom. Motivation among academic enablers refers to the student’s willingness to participate and persist in learning until having satisfied a goal (DiPerna, 2006). Motivation addresses working with effort. Furthermore, motivation influences academic engagement and study
skills (DiPerna et al., 2002). Academic engagement refers to attentive, active participation in class instruction, specifically when the student listens to and respects the teacher along with following the teacher’s directions (DiPerna & Elliot, 2000; Greenwood, Horton, & Utley, 2002). Study skills involve self-regulated learning and include such behaviors as doing seatwork as directed and making assistance needs known in an appropriate manner (DiPerna & Elliot, 2000). The results of the Pennefather and Smolkowski (2015) study indicate a single factor structure for the latent variable of teachability.

**Student Risk Screening Scale.** The *Student Risk Screening Scale* (SRSS) measures the construct of externalizing problems (Drummond, 1994). Six recent studies by Lane and colleagues have established Cronbach’s alpha (α = .78 to .87), indicating good internal consistency (Ennis, Lane, & Oakes, 2012; Lane, Little, et al., 2009; Lane, Kalberg, Lambert, Crnobori, & Bruhn, 2010; Lane, Menzies, et al., 2012; Menzies & Lane, 2012; Oakes et al., 2010). Furthermore, the range of test-retest reliability shows stability between fall and winter (r = .73 to .79), winter and spring (r = .74 to .86), and fall and spring (r = .68 to .77). Whereas the SRSS discriminates students with poor social skills based on the *Social Skills Improvement System- Performance Screening Guide* (r = -0.50 to -0.63, p < .0001; Lane, Richards-Tutor, Oakes, & Connor, 2013), the SRSS aligns with scores from the *Strengths and Difficulties Questionnaire* with (r = .71 to .80, p < .0001; Ennis et al., 2012).

Externalizing problems involve either antisocial or disrespect behavior patterns (Gresham, 2015). The antisocial behaviors violate social norms through aggression and relational aggression. They include bullying, fighting, stealing, cheating, and lying. The
disrespectful behaviors instigate conflict with peers and noncompliance with authority figures. They include arguing, complaining, provoking others, defiance, temper outbursts, and resentfulness. Externalizing problems often develop into oppositional defiant disorder, conduct disorder, or substance abuse disorder.

**Student Internalizing Behavior Screener.** The *Student Internalizing Behavior Screener* (SIBS) measures the level of manifest signs of distress from internalizing symptoms (Cook et al., 2011). Cronbach’s alpha for internal consistency reached $\alpha = .81$ in the fall and $\alpha = .79$ in the winter. The test-retest correlation coefficient between fall and winter administrations was $r = .74$. The SIBS demonstrated convergent validity ($r = .82$) with the internalizing scale of the *Child Behavior Checklist - Teacher Report Form* (Achenbach, 1991).

Internalizing problems include distressing symptoms associated with depression, anxiety, and social withdrawal. Depressive syndromes may involve excessive sadness or frequently crying; anxious syndromes may involve excessive worrying or fearfulness; and social withdrawal entails the avoidance of social interactions and support. Excessive control of the emotional and cognitive states arising from these symptoms leads to internalizing disorders (Cicchetti & Toth, 1991). A common mediator of internalizing symptoms is avoidance of unpleasant or unwanted thoughts, feelings, sensations, and events (Hayes et al., 1996). Whereas adults readily observe externalizing problems as outwardly directed violations of social norms, they often have difficulty recognizing self-directed internalizing problems occurring within students.
Procedures for Data Collection

The winter administration occurred between February 22nd and March 9th of 2016. Schools administered the screeners during regularly scheduled staff meetings either before or after school. The screeners were administered online in the following order: (a) ESBA, (b) SRSS, and (c) SIBS. See Appendix B for the online directions for completing the screeners.

A day prior to the administration, the director of implementation from the University of Washington (UW) project explained to the teachers at each of the schools the purpose and the process of the universal mental health screening to the teachers at each of the schools. The 30-minute training session engaged all the available teaching staff. The following day, the university staff and site leadership addressed additional questions and restated the purpose and process of the screeners in the context of their multi-tiered systems of support. Additionally, the information technology specialist from the school district assisted teachers with accessing the online screeners.

During the administration session, teachers were instructed to log onto the district-wide student information system (Performance Plus by Sunguard) and verify their class roster (See Appendix C for a screenshot of the online screeners). Teachers then completed the ESBA in the same room. After nearly 20 minutes, the school principal or UW implementation director prompted the teachers to complete the SRSS. Teachers were prompted in another 20 minutes to complete the SIBS last. Finally, teachers were advised that, had they not finished during the 60-minute session, they were to do so independently. The UW staff monitored the completion of the screeners and prompted
administrators to follow through with teachers who had been absent on the day of administration.

Since the screeners were administered through the student information system, the data were stored in the district data warehouse. Parental consent for the screeners was not required because every student is considered during universal screening, much as with vision and hearing screenings. The data were not used for special education services. The school district provided the de-identified data in accordance with their district policy and the Research Review Board.

Data Analysis

This study used multigroup confirmatory factor analysis to examine measurement equivalence of the latent variables across racial and ethnic groups. Each of the screeners was analyzed separately as a single-factor model across five racial and ethnic groups. African American, Asian American or Pacific Islander, Latino Hispanic, and multiracial groups were compared to the European American group on the factor structure, factor loadings, and thresholds for the latent variables. This study used the weighted least squares with means and variance adjustment (WLSMV) for the estimation of the nested models because the ordered categories of the rating scales did not meet the assumptions of continuous variables for maximum likelihood methods (DiStefano & Morgan, 2014; Finney & DiStefano, 2006; Lei, 2009; Rhemtulla, Brosseau-Liard, & Savalei, 2012). Please refer to Appendix D for the comparison of WLSMV to other estimation methods.
**Measured variables.** Both the independent and dependent variables in this study were categorical. The independent variable was the racial and ethnic group for each student. The dependent variables were the ordinal ratings from the screeners for the latent variables of teachability, externalizing problems, and internalizing problems.

**Data analysis procedures.** The initial data analysis consisted of evaluating the quality of data and the assumptions inherent in the statistical analyses. The items from each screener were examined for multivariate normality. The Mardia’s test for multivariate normality measured both kurtosis and skewness of the four categories of the ordinal scales, i.e., *never, occasionally, sometimes,* and *frequently.* The distribution of the scores likely exhibited extreme skewness and kurtosis as expected, due to the potential base rate of 20% of students presenting a disorder at some point during the school year (Perou et al., 2013). In other words, the majority of students obtained low scores on the measures of externalizing and internalizing behavior problems, presenting a positive skew and elevated kurtosis. During each step of testing the nested models, each indicator was examined for Heywood cases in which the standardized loading was greater than 1.0 and the error variance was negative.

**Steps for testing for measurement invariance with confirmatory factor analysis.** The data analysis procedures for measurement invariance across multiple racial and ethnic groups using confirmatory factor analysis (CFA) from structural equation modeling addressed each of the research questions. Brown (2015) has recommended the following steps for multi-group CFA for measurement invariance:

1. Test the CFA model separately for each group.
2. Test whether the configuration of the indicators is equivalent across all of the groups.

3. Test whether the factor loadings on each indicator are equivalent across all of groups.

4. Test whether the thresholds as an intercept for each indicator are equivalent across all of the groups.

These steps were conducted separately for each latent variable in the order they were administered—specifically, teachability, externalizing problems, and internalizing problems.

Step one in the CFA procedure involved identifying the baseline model for the factor structure of each screener on each group separately. The baseline model for an individual group established the best-fitting model in which the latent variable loads substantially onto the indicators. If an indicator held a trivial loading for a specific group, then the baseline model failed to support the inclusion of that individual group in the multigroup analysis of measurement invariance. Thus, the testing sequence stops for any model of a group that exhibits major differences in the factor structure and a poor model fit. The models that demonstrated an adequate fit along with nontrivial loadings on the indicators can stand on their own. The multigroup CFA included only groups with an adequate model of the factor structure of latent variable for each screener.

**Research question 1.** Does the screener measure the latent construct of interest with the equivalent factor structure consisting of nontrivial indicators across each of the racial and ethnic groups compared to the index group?
Step two in the CFA procedure addressed the first research question for each screener by testing the degree to which the particular screener simultaneously shared the same meaning in the latent construct across each of the racial and ethnic groups. In step two, the procedure compared the racial and ethnic groups to the index group for the equivalent factor structure consisting of non-trivial indicators. Step two tested simultaneously all of the separate groups that established an adequate baseline model in step one. In step two, none of the parameters except the reference indicator was constrained; that is, they were allowed to vary. Any group whose model did not share the equivalent meaning in the latent variable for the screener did not continue to subsequent steps in the analysis. Simultaneous configural invariance across racial and ethnic groups established the prerequisite for step three.

**Research question 2.** Do the ordinal descriptors of *never, occasionally, sometimes, and frequently* demonstrate measurement equivalence in the metric unit of frequency on for latent construct of interest for the screener across each of the racial and ethnic groups compared to the European American index group?

Step three in the CFA procedure addressed the second research question for each screener by testing for metric unit equivalence of the descriptors of *never, occasionally, sometimes, and frequently* on latent constructs across each of racial and ethnic groups relative to the European American index group. The frequency descriptors represent ordinal data in which the magnitude of the difference between descriptors may have been inconsistent. The difference between *occasionally* and *sometimes* does not have an established universal meaning. Therefore, step three tested whether teachers applied the
descriptors of *never*, *occasionally*, *sometimes*, and *frequently* with the same meaning across racial and ethnic groups.

Metric equivalence means that a change in one unit on the obtained score of the screener in one racial or ethnic group holds the same value for the other groups as well. In order to establish equivalence in the metric unit, each of the groups must share a value similar to the index group’s factor loading. Thus, models for the racial and ethnic groups had their factor loadings constrained to the values of the factor loadings from the European American group. For each screener, the model with the equality constrained factor loadings in step three was compared to the model with unconstrained parameters from step two. If the model in step three demonstrated goodness of fit with the model from step two, then invariant groups that shared the same unit value on the score, indicating that the ratings of *never*, *occasionally*, *sometimes*, and *frequently* on latent constructs, held the same meaning across each group. Metric invariance established the condition for step four.

**Research question 3.** Does the descriptor, *never*, establish measurement equivalence on the initial threshold of the scale of the construct of interest for the screener across each of the racial and ethnic groups compared to the index group?

Step four addressed research question three for each screener by testing whether the descriptor *never* shared the same threshold across each of racial and ethnic groups relative to the European American index group. The threshold of an ordinal variable fixes the origin for the first category in a series of ordered categories, functioning much like the intercept of a continuous variable by setting the starting point for the scale of the measurement unit.
In step four, both the factor loading and factor thresholds were constrained to the values from the index group. The model with constrained factor loadings and factor thresholds in step four were compared simultaneously with the model with constrained factor loadings only from step three. If the model from step four demonstrated goodness of fit with the model from step three, then invariant groups shared the equivalent value on the scale of measurement and the ordinal descriptor *never* held the same meaning of the across racial and ethnic groups.

**Evaluation of the goodness of fit for models.** Goodness of fit statistics quantify the degree to which the implied baseline model reproduces the observed model data set. The goodness-of-fit statistics in this study took into account the large sample size using ordinal data (Brosseau-Liard, Savalei, & Li, 2012). Evaluating the goodness of fit between the models begins with the chi squared test of exact fit followed by a test of close fit based on comparison of the residuals and lastly the incremental tests of approximate fit based on chi squared distribution. These different fit indices quantify the absolute fit, the fit adjusted for model complexity and parsimony, and the fit compared to a null model (Chen, 2007).

The $\chi^2$ exact-fit test tests the null hypothesis that the implied baseline model is different from the observed model. However, the $\chi^2$ exact fit test is overly sensitive to large samples and non-normal data. Specifically, large sample sizes ($N > 400$) often lead to small yet statistically significant differences between the baseline and observed models, resulting in rejecting baseline and observed models being nearly the same, though not exactly same (Bentler & Bonnett, 1980; Brown, 2015; Kline, 2016).
The Root Mean Square Error of Approximation (RMSEA; Steiger, 1990) provides an absolute fit index for the degree to which the discrepancy exists between the baseline model and observed model based. Kline (2016) recommends Browne and Cudeck’s (1993) cut values on the RMSEA with below 0.05 indicating a close fit, values above .05 supporting the not-close but still fair fit hypothesis, and values above .10 suggesting a poor fit. A large sample may improve the precision of the RMSEA to identify models that fail the exact test based on the $\chi^2$ test, yet may still be close to the fit based on the $p$ values.

The incremental fit test of the Comparative Fit Index (CFI; Bentler, 1990) evaluates the fit of the baseline model nested within the observed model. Brown (2015) has stated that CFI values above .90 suggest an acceptable fit and above .95 a good fit. The CFI compares the amount of difference between the implied model and the close fit of the observed model. A CFI of 0.95 indicates that the observed model appears 95% better fit than the baseline model. The Tucker Lewis Fit Index (TLI; Tucker & Lewis, 1973) was not included in this study because of the high correlation with the CFI (Kenny, 2016).

**Evaluation of measurement invariance in nested models.** Measurement invariance involves comparing a nested model with more constraints to the parent model with fewer constraints to determine whether the difference between the models is significant. As the parameters are freed to vary and the degrees of freedom increase, the value of the CFI is likely to differ between the nested and parent model. The magnitude of change in the CFI ($\Delta$ CFI = CFI\textsubscript{parent} - CFI\textsubscript{nested}) above 0.01 indicates that models are different (Chen, 2007). Furthermore, a positive difference in $\Delta$ CFI means that the more
constrained model fits worse than the unconstrained model, indicating that the groups demonstrate non-invariance at the level of breakdown in the models (Cheung & Rensvold, 2002). Specifically, Δ CFI may indicate a breakdown between configural and metric models based on the non-invariant loadings or between the metric and scalar models based on non-invariant thresholds. As the number of freed parameters increases, the model becomes more complex, thereby making the fit between the null model and the observed model more difficult to achieve (Kang, McNeish, & Hancock, 2016).
Chapter Four: Results

The purpose of this study was to test the degree to which screening instruments demonstrate both equivalent meaning and equivalent measurement properties across five racial and ethnic groups. The screening instruments include the Elementary Social Behavior Assessment (ESBA; Pennefather & Smolkowski, 2015), the Student Risk Screening Scale (SRSS; Drummond, 1994), and the Student Internalizing Behavior Screener (SIBS; Cook et al., 2011). This chapter begins with a brief description of the sample and the data. Before the data were analyzed with respect to the research questions, the baseline models for each screener were established, beginning with the omnibus model that included every participant, in addition to models for each racial and ethnic group separately. Once the baseline models demonstrated an adequate fit, the multigroup confirmatory factor analysis for measurement invariance addressed the first research question by simultaneously testing each group for the pattern of meaningful loadings on the latent variable. The second research question compared the model fit from configural invariance with the nested model for the test of metric invariance in which the factor loadings are constrained. The third research question compared the fit from the metric model with the test for scalar invariance in which both the thresholds and factor loadings are constrained.

Description of the Sample and Data

This study used a multistage sampling method. From the initial sample frame of 36 elementary schools, 10 schools satisfied the criteria for selection by meeting the 80-80 rule on the School-wide Evaluation Tool (SET; Horner et al., 2004; Sugai et al., 2001),
completing the school-wide PBIS Tiered Fidelity Inventory (TFI; Algozzine et al., 2014), and presenting their school-wide action plans for addressing mental health needs of children. The primary sample unit of the 10 selected schools established the secondary sampling unit in which teachers rated all of the students assigned to general education classrooms using the three mental health screeners.

**Demographic data.** The original data on 4012 students were reduced to 3965 students. The data set excluded American Indian students and students with unspecified racial or ethnic data due to small sampling. The racial and ethnic composition of the final sample included Asian American/Pacific Islander (13.6%), African American (18.4%), Latino Hispanic (25.4%), multiracial (11.1%), and European American (31.5%) students. The distribution of students in kindergarten through fifth-grade represented each grade level with a range of 15.8% to 17.6%. Additionally, 51.4% of the sample was male and the remaining 48.6% female.

**Power.** A common approach for power analysis in confirmatory factor analysis involves the Root Mean Square Error of Approximation (RMSEA). The degrees of freedom, along with the sample size, set the critical value of the chi-square distribution for a given alpha significance level ($\alpha = .05$) for each model (MacCallum et al., 1996). The models for the SRSS and SIBS had the fewest degrees of freedom (70) in the configural model and the group of multiracial students has the fewest students (441). For the model with the fewest degrees of freedom and the smallest sample size, the resulting power equals .999. Furthermore, when the analysis included all five groups, the power remained at .999. Thus, the sample sizes for analyzing specific racial and ethnic groups
as well as comparing racial and ethnic groups sufficiently established the power to detect differences in their models.

**Missing data.** The percentage of missing data fell below 1% for both the ESBA and the SIBS and approximated 2% for the SRSS. The *lavaan* statistical application uses listwise deletion for missing data for ordinal ratings (Rosseel, 2012; Rosseel et al., 2017). Therefore, the analysis of each screener has a different quantity of participants.

Although teachers completed every instrument that they started, some teachers did not start every instrument. The groupings for missing data occurred in six of the ten schools. The most frequent pattern of missing data came from a specific grade level within a given school. For example, one teacher from a grade level with three teachers may not have completed a particular screener for every student, resulting in the missing data groups within the grade level for that school. The 43 missing ESBA screeners occurred in one fifth-grade grouping (8 missing) and two second-grade groupings from different schools (21 and 12 missing). The ESBA also had two randomly missed screeners. The 73 missing SRSS included one fifth-grade grouping (eight missing), two groupings from the same school with 24 missing in the fourth grade and 12 missing from the second grade, and a third grade grouping (26 missing). The SRSS also had randomly missed screeners. The SIBS had 35 missing screeners with a third grade grouping (23 missing) and a fifth grade grouping (eight missing). There were four randomly missing SIBS in kindergarten.

**Data normality.** Ordinal data often do not meet the assumptions of normality (Raykov & Marcoulides, 2015). Ratings with four ordered categories (*never, occasionally, sometimes, frequently*) do not perform as well as ordinal data with more
categories with regard to normality (Finney & DiStefano, 2006; Rhemtulla et al., 2012). Severe skewness occurs with values over three and kurtosis with values between 8 and 20 (Kline, 2016).

Table 5
Univariate Skewness and Kurtosis on the ESBA, SRSS, and SIBS

(a) ESBA

<table>
<thead>
<tr>
<th>Item</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>T5</th>
<th>T6</th>
<th>T7</th>
<th>T8</th>
<th>T9</th>
<th>T10</th>
<th>T11</th>
<th>T12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skewness</td>
<td>-1.74</td>
<td>-1.54</td>
<td>-1.50</td>
<td>-1.55</td>
<td>-1.53</td>
<td>-1.61</td>
<td>-1.43</td>
<td>-1.77</td>
<td>-1.53</td>
<td>-2.46</td>
<td>-1.87</td>
<td>-1.53</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>5.60</td>
<td>4.93</td>
<td>4.56</td>
<td>4.78</td>
<td>4.65</td>
<td>5.15</td>
<td>4.41</td>
<td>5.65</td>
<td>4.55</td>
<td>8.98</td>
<td>6.37</td>
<td>4.71</td>
</tr>
<tr>
<td>All ps&lt;.001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(b) SRSS

<table>
<thead>
<tr>
<th>Item</th>
<th>E1</th>
<th>E2</th>
<th>E3</th>
<th>E4</th>
<th>E5</th>
<th>E6</th>
<th>E7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skewness</td>
<td>3.76</td>
<td>1.67</td>
<td>1.20</td>
<td>2.09</td>
<td>0.77</td>
<td>1.51</td>
<td>2.30</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>17.52</td>
<td>4.69</td>
<td>2.97+</td>
<td>6.38</td>
<td>2.01</td>
<td>3.95</td>
<td>7.28</td>
</tr>
<tr>
<td>All ps &lt; .001, except + E3 p = 0.77</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(c) SIBS

<table>
<thead>
<tr>
<th>Item</th>
<th>I1</th>
<th>I2</th>
<th>I3</th>
<th>I4</th>
<th>I5</th>
<th>I6</th>
<th>I7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skewness</td>
<td>2.21</td>
<td>3.00</td>
<td>2.14</td>
<td>0.76</td>
<td>2.29</td>
<td>1.70</td>
<td>2.31</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>7.04</td>
<td>11.98</td>
<td>6.72</td>
<td>1.98</td>
<td>7.44</td>
<td>4.84</td>
<td>7.45</td>
</tr>
<tr>
<td>All ps &lt; .001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: **Bold** indicates severe univariate non-normality.

Table 5 presents the level of univariate skewness and kurtosis for the ESBA in (a), the SRSS in (b), and the SIBS in (c). On the ESBA, 8% of the items exhibit severe univariate skewness and kurtosis. On the SRSS, 43% of the items show severe skewness and 29% severe kurtosis. On the SIBS, 71% of the items show severe skewness and 57% severe kurtosis. Elevated univariate skewness and kurtosis contributes to multivariate
non-normality. The tests of multivariate normality confirm severe skewness and kurtosis for each of the screeners. Mardia’s test for multivariate skewness shows 36.86 on the ESBA, 34.62 on the SRSS, and 32.01 on the SIBS (\( p < .001 \)). Mardia’s test for multivariate kurtosis shows 349.88 on the ESBA, 129.28 on the SRSS, and 134.42 on the SIBS (\( p < .001 \)). These elevated levels of multivariate non-normality indicate that the data does not meet the assumptions of normality. There are no multivariate outliers.

**Results**

Testing the CFA model for each screener with an omnibus CFA and separate CFA models for each racial and ethnic group preceded testing for measurement invariance. Acceptable CFA models with the RMSEA < .08 and the CFI > .90 justified further analysis for measurement invariance based on recommendations from Brown (2015) and Kline (2016). Good models reach the minimum threshold for two fit measures; marginal models meet cut score on any one of the fit measures; and poor models fail to pass the minimum threshold on any measure. The model \( \chi^2 \) test was not used to evaluate the goodness of fit due to the large sample size and the degree of multivariate non-normality. In addition to the results presented below, the polychoric correlation matrix for the indicators appears in appendix F and additional post-hoc analyses using different estimators appear in the appendix D.

**Elementary Social Behavior Assessment.** The omnibus CFA for the ESBA (\( \chi^2 \) (54) = 2338.967, \( p < .001 \); RMSEA = 0.104, CFI = 0.624) indicated that problems would potentially occur with separate unconstrained models for each racial and ethnic group (Table 6). The CFI and the RMSEA exceeded criteria for each of the groups as well.
The separate CFA models demonstrated poor fit across European American ($\chi^2 (54) = 840.335, p < .001, \text{RMSEA} = 0.102, \text{CFI} = 0.596$); African American ($\chi^2 (54) = 536.862, p < .001, \text{RMSEA} = 0.111, \text{CFI} = 0.775$); Asian American/Pacific Islander ($\chi^2 (54) = 217.431, p < .001, \text{RMSEA} = 0.075, \text{CFI} = 0.561$); Latino Hispanic ($\chi^2 (54) = 504.460, p < .001, \text{RMSEA} = 0.091, \text{CFI} = 0.567$); and multiracial groups of students ($\chi^2 (54) = 289.935, p < .001, \text{RMSEA} = 0.100, \text{CFI} = 0.730$). The unconstrained models of ESBA demonstrated the need for respecification.

The conceptual basis for the respecification aligns the skills that (a) strengthen the student-teacher relationship, (b) build friendships, and (c) promote school adjustment and academic success (Walker et al., 2015). Respecifying the measurement model was needed to allow for error variances to covary due to the similarity in the performance of the indicators. These correlated errors between indicators conform to the single factor model from the exploratory factor analysis (Pennefather & Smolkowski, 2015).

The first step toward analyzing the failure of the initial ESBA model started with examining the residuals between the implied model and the observed for areas of the weaknesses. The residuals matrix of the polychoric correlations provides a standardized estimate of the relationships between the items based on ordinal data. Polychoric correlation residuals above .10 indicate a mismatch between the two models. The positive elevated polychoric residuals suggested that the implied model would underpredict their observed association (see bold in Appendix G).

Table 6

*Initial Unconstrained Model on the Elementary Social Behavior Assessment*
<table>
<thead>
<tr>
<th>Unconstrained Models</th>
<th>$\chi^2$ (df)</th>
<th>RMSEA (90% CI)</th>
<th>CFI</th>
<th>Description of model fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Groups (N = 3937)</td>
<td>2338.967 (54)</td>
<td>0.104, 0.107</td>
<td>0.624</td>
<td>Poor</td>
</tr>
<tr>
<td>European American (N = 1237)</td>
<td>840.335 (54)</td>
<td>0.102, 0.109</td>
<td>0.596</td>
<td>Poor</td>
</tr>
<tr>
<td>African American (N = 722)</td>
<td>536.862 (54)</td>
<td>0.111, 0.120</td>
<td>0.775</td>
<td>Poor</td>
</tr>
<tr>
<td>Asian American/Pacific Islander (N = 536)</td>
<td>217.431 (54)</td>
<td>0.075, 0.086</td>
<td>0.561</td>
<td>Poor</td>
</tr>
<tr>
<td>Latino Hispanic (N = 1004)</td>
<td>504.460 (54)</td>
<td>0.091, 0.099</td>
<td>0.567</td>
<td>Poor</td>
</tr>
<tr>
<td>Multi-Racial (N = 438)</td>
<td>289.935 (54)</td>
<td>0.100, 0.111</td>
<td>0.730</td>
<td>Poor</td>
</tr>
</tbody>
</table>

The pairs of polychoric correlations with elevated positive values suggested that the items were difficult to distinguish. These items may have captured the same concepts in which the lack of precision in the wording in one item may influence the understanding of the other item in the pair. For example, a teacher may not distinguish between a student who works with effort (T3) and a student who does seatwork as directed (T4). The items pairs of T3-T4, T10-T11, T10-T12, and T11-T12 follow similar patterns of overlapping meaning. The T3-T4 pair has the highest residual of 0.238. The
next elevated residual (0.126) correlates *gets along with peers* (T11) with *resolves peer conflict* (T12). The remaining elevated residuals (0.102) pairs *can have normal conversations without becoming hostile* (T10) with both T11 and T12. Getting along with peers, resolving peer conflict, and engaging in conversations without hostility may be difficult to distinguish on a scale of frequency. The re-test of the first respecified ESBA model involved setting the residuals to covary between item pairs with positive residuals above .10 and improved the CFI from 0.624 to 0.760 (Table 7).

The modification indices for the initial ESBA model showed that 19 of 66 pairs of the present z scores were above the Wald of 3.84, indicating that potentially 29% of the model could include additional covariances in the residuals. Specifically, eight residuals with modification indices above 6.75 made conceptual sense to set their error variances to covary (T1-T2, T2-T3, T2-T4, T3-T5, T4-T5, T9-T10, T9-T11, and T9-T12). Adding to the previous T3-T4 covariance, a teacher may perceive the student who can *make assistance needs known in appropriate manner* (T5) as also *working with effort* (T3-T5 modification index 21.289) and *doing seatwork as directed* (T4-T5 modification index 21.836). Furthermore, a student who is *working with effort* (T3) also *follows the teacher's directions* (T2-T3 modification index 8.291) and such a student *listens to and respects the teacher* (T1-T2 modification index 9.254). Similarly, the student who *works out strong feelings appropriately* (T9) also *can have normal conversations without becoming hostile* (T9-T10 modification index 13.947), *gets along with peers* (T9-T11 modification index 12.454) and *resolves peer conflict* (T9-T12 modification index 22.298). The final respecified omnibus ESBA model with 12 pairs of covarying residuals improved the CFI to 0.931 (Table 7).
Table 7
Respecified Unconstrained Models with Covarying Residuals on the Elementary Social Behavior Assessment

<table>
<thead>
<tr>
<th>Models</th>
<th>$\chi^2$ (df)</th>
<th>RMSEA (90% CI)</th>
<th>CFI</th>
<th>Description of model fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retest 1 with covarying residuals for T3-T4, T10-T11, T10-T12, and T11-T12</td>
<td>1510.895 (50)</td>
<td>0.086 0.082 0.090</td>
<td>0.760</td>
<td>Poor</td>
</tr>
<tr>
<td>Retest 2 with covarying residuals from restest 1 and for T1-T2, T2-T3, T2-T4, T3-T5, T4-T5, T9-T10, T9-T11, and T9-T12</td>
<td>461.184 (42)</td>
<td>0.050 0.046 0.055</td>
<td>0.931</td>
<td>Marginal</td>
</tr>
<tr>
<td>European American</td>
<td>206.091 (42)</td>
<td>0.056 0.049 0.064</td>
<td>0.916</td>
<td>Marginal</td>
</tr>
<tr>
<td>African American</td>
<td>136.474 (42)</td>
<td>0.056 0.046 0.066</td>
<td>0.956</td>
<td>Good</td>
</tr>
<tr>
<td>Latino Hispanic</td>
<td>116.907 (42)</td>
<td>0.042 0.033, 0.051</td>
<td>0.928</td>
<td>Marginal</td>
</tr>
<tr>
<td>Asian American/Pacific Islander</td>
<td>54.437 (42)</td>
<td>0.024 0.000, 0.040</td>
<td>0.967</td>
<td>Good</td>
</tr>
<tr>
<td>Multi-Racial</td>
<td>107.525 (42)</td>
<td>0.060 0.046, 0.074</td>
<td>0.925</td>
<td>Marginal</td>
</tr>
</tbody>
</table>

The CFA models for each of the racial and ethnic groups reached adequate levels of fit (Table 7). The models for African American students $\chi^2 (42) = 136.474, p < .001,$
RMSEA = 0.056, CFI = 0.956) and for Asian American/Pacific Islander ($\chi^2$ (42) = 54.437, $p < .001$, RMSEA = 0.024, CFI = 0.967) made a good fit. The models for European American ($\chi^2$ (42) = 206.091, $p < .001$, RMSEA = 0.056, CFI = 0.916); Latino Hispanic ($\chi^2$ (42) = 116.907, $p < .001$, RMSEA = 0.042, CFI = 0.928); and multiracial groups of students ($\chi^2$ (42) = 107.525, $p < .001$, RMSEA = 0.060, CFI = 0.925) reached a marginal fit across the indices.

Figure 5. Unstandardized Parameter Estimates for the respecified omnibus ESBA. Note: the 12 covariances between the residuals appear stacked between indicators.

The respecified model of the ESBA set the foundation for testing for measurement invariance (Figure 5). The factor loadings appear in the scale of the screener to assist with interpretation. Specifically, all of the factor loadings round to the nearest unit value of one, suggesting that a one-unit change in frequency of *listens to and respects the teacher* nearly equates to a one-unit change in the other indicators. The interpretability of the ratings will require validation with another sample to prevent overreliance on chance in such an extremely large sample (MacCallum, Roznowski, & Necowitz, 1992). Nonetheless, all of the standardized factor loadings exceeded .70
(range .71 to .89). Standardized factor loadings can be interpreted much like a regression coefficient.

**Student Risk Screening Scale.** The omnibus CFA for the SRSS ($\chi^2 (14) = 194.338, p < .001; \text{RMSEA} = 0.058, \text{CFI} = 0.948$) warrants further analysis of unconstrained models for each racial and ethnic group. The models for African American students ($\chi^2 (14) = 59.801, p < .001, \text{RMSEA} = 0.068, \text{CFI} = 0.961$) and Latino Hispanic students ($\chi^2 (14) = 32.004, p < .001, \text{RMSEA} = 0.036, \text{CFI} = 0.965$); multiracial students ($\chi^2 (14) = 46.433, p < .001, \text{RMSEA} = 0.073, \text{CFI} = 0.940$) and European American students ($\chi^2 (14) = 87.888, p < .001, \text{RMSEA} = 0.065, \text{CFI} = 0.934$) met the criteria for a good fit. The fit indexes for the Asian American/Pacific Islander students ($\chi^2 (14) = 28.218, p < .001, \text{RMSEA} = 0.044, \text{CFI} = 0.927$) indicated a marginal fit for their model. The CFA models of the SRSS for each racial and ethnic group indicated an adequate fit for testing for measurement invariance (Table 8).

![Diagram](image)

*Figure 6. Unstandardized Parameter Estimates for the SRSS.*

The CFA for the SRSS for each racial and ethnic group presented an adequate unconstrained model for addressing the research questions. Figure 6 illustrates the model
in the scale of the SRSS to assist with interpretation. Item three set the reference indicator so that an unstandardized unit change in the frequency of *behavior problems* may also involve an unstandardized unit change in the other indicators, except for *stealing*. Furthermore, the standardized factor loadings appear strong (> .7) for items E2, E3, E6, and E7. The standardized factor loadings for E4 (.69) falls just short of a strong loading. The standardized factor loadings for E1 (.56) and E5 (.42) contribute adequately to the model’s purpose of screening.

Table 8
*Unconstrained Models for the Student Risk Screening Scale*

<table>
<thead>
<tr>
<th>Models</th>
<th>$\chi^2$ (df)</th>
<th>RMSEA (90% CI)</th>
<th>CFI</th>
<th>Description of model fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Groups (N = 3896)</td>
<td>194.338 (14)</td>
<td>0.058</td>
<td>0.948</td>
<td>Good</td>
</tr>
<tr>
<td>European American (N = 1233)</td>
<td>87.888 (14)</td>
<td>0.065</td>
<td>0.934</td>
<td>Good</td>
</tr>
<tr>
<td>African American (N = 716)</td>
<td>59.801 (14)</td>
<td>0.068</td>
<td>0.961</td>
<td>Good</td>
</tr>
<tr>
<td>Latino Hispanic (N = 987)</td>
<td>32.004 (14)</td>
<td>0.036</td>
<td>0.965</td>
<td>Good</td>
</tr>
<tr>
<td>Asian American/Pacific Islander (N = 530)</td>
<td>28.218 (14)</td>
<td>0.044</td>
<td>0.927</td>
<td>Marginal</td>
</tr>
<tr>
<td>Multi-Racial (N = 430)</td>
<td>46.433 (14)</td>
<td>0.073</td>
<td>0.940</td>
<td>Good</td>
</tr>
</tbody>
</table>
Student Internalizing Behavior Scale. The adequacy of the omnibus CFA for the SIBS ($\chi^2(14) = 127.926, p < .001; \text{RMSEA} = 0.045, \text{CFI} = 0.950$) established the basis for analyzing the unconstrained models for each racial and ethnic group. The fit indexes for each racial and ethnic group included the following: the models for European American students ($\chi^2(14) = 45.221, p < .001, \text{RMSEA} = 0.065, \text{CFI} = 0.961$) and multiracial students ($\chi^2(14) = 24.525, p < .001, \text{RMSEA} = 0.042, \text{CFI} = 0.972$) demonstrated good fit across the three indices. The fit indices for African American ($\chi^2(14) = 50.030, p < .001, \text{RMSEA} = 0.060, \text{CFI} = 0.922$); Asian American/Pacific Islander ($\chi^2(14) = 26.069, p < .001, \text{RMSEA} = 0.040, \text{CFI} = 0.907$); and Latino Hispanic ($\chi^2(14) = 46.904, p < .001, \text{RMSEA} = 0.049, \text{CFI} = 0.932$) reached marginal fit due to lower TLI values. The CFA models of the SIBS for each racial and ethnic group demonstrated adequacy for testing for measurement invariance (Table 9).

The CFA model for the SIBS for each racial and ethnic group presented an adequate unconstrained model for addressing the research questions. Figure 7 depicts the model in the scale of the SIBS. All of the unstandardized factor loadings round to the nearest unit value of one, suggesting that a one-unit change in frequency of nervous and fearful nearly equates to a one-unit change in the other indicators. The standardized factor loadings for Items I5 and I6 appear strong (>.70). Items I1, I2, and I3 show moderate standardized factor loadings (>.50). Items I4 and I7 with standardized loadings just above .40 adequately support the model for the purpose of screening.
Table 9
Unconstrained Models on the Student Internalizing Behavior Scale

<table>
<thead>
<tr>
<th>Description of model fit</th>
<th>RMSEA (90% CI)</th>
<th>CFI</th>
<th>$\chi^2$ (df)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Groups (N = 3933)</td>
<td></td>
<td></td>
<td>127.926 (14)</td>
</tr>
<tr>
<td>European American (N = 1241)</td>
<td></td>
<td></td>
<td>45.221 (14)</td>
</tr>
<tr>
<td>African American (N = 722)</td>
<td></td>
<td></td>
<td>50.030 (14)</td>
</tr>
<tr>
<td>Latino Hispanic (N = 999)</td>
<td></td>
<td></td>
<td>46.904 (14)</td>
</tr>
<tr>
<td>Asian American/Pacific Islander (N = 534)</td>
<td></td>
<td></td>
<td>26.069 (14)</td>
</tr>
<tr>
<td>Multi-Racial (N = 437)</td>
<td></td>
<td></td>
<td>24.525 (14)</td>
</tr>
</tbody>
</table>

Figure 7. Unstandardized parameter estimates for the SIBS.
**Preliminary results to the research questions.** The preceding omnibus CFA and separate CFA models for each racial and ethnic group exhibited adequate goodness of fit to permit further tests for measurement invariance. The preliminary results address each of the basic research questions. The results of the tests measurement invariance appear in Table 10 for the ESBA, Table 11 for the SRSS, and Table 12 for the SIBS.

**Research question 1.** Does the screener measure the latent construct of interest with the equivalent factor structure consisting of nontrivial indicators across each of the racial and ethnic groups relative to the European American index group?

The ESBA, SRSS, and SIBS measure the constructs of teachability, externalizing problems, and internalizing problems with the equivalent factor structure with the same pattern of constrained and fixed parameters across each racial and ethnic group. The respecified ESBA demonstrated marginal fit ($\chi^2 (250) = 651.080, p < .001$, RMSEA = 0.052, CFI = 0.928) on the test for configural invariance. The tests for unconstrained thresholds and loadings simultaneously comparing each racial and ethnic group on the SRSS ($\chi^2 (70) = 256.245, p < .001$, RMSEA = 0.058, CFI = 0.947) and on the SIBS ($\chi^2 (70) = 190.867, p < .001$, RMSEA = 0.047, CFI = 0.945) indicated good fit on configural invariance. The respecified ESBA, the SRSS, and the SIBS established meaningful factor loadings for each racial and ethnic group compared to the European American group.

**Research question 2.** Do the ordinal descriptors of *never*, *occasionally*, *sometimes*, and *frequently* demonstrate measurement equivalence in the metric unit of frequency on latent construct of interest for the screener across each of the racial and ethnic groups relative to the European American index group?
In order to compare the metric model nested within the configural model, the factor loadings must be constrained. The change in the CFI from configural model to the constrained loadings model was less .01 for the ESBA, SRSS and the SIBS. Instead of diminishing the goodness of the fit, the multigroup CFA with constrained factor loadings improved the CFI from the configural model to the metric model by .039 on the ESBA, .024 on the SRSS, and .033 on the SIBS. These improvements in the fit due to the constrained loadings indicated that the frequency ratings shared the same metric unit across racial and ethnic groups. Specifically, the ordinal descriptors of *never*, *occasionally*, *sometimes*, and *frequently* shared the same meaning for frequency across each of the racial and ethnic groups on the ESBA ($\chi^2(254) = 454.715, p < .001, \text{RMSEA} = 0.032, \text{CFI} = 0.967$), the SRSS ($\chi^2(94) = 197.078, p < .001, \text{RMSEA} = 0.038, \text{CFI} = 0.971$), and the SIBS ($\chi^2(94) = 141.770, p < .001, \text{RMSEA} = 0.025, \text{CFI} = 0.978$). Each screener demonstrated metric invariance.

**Research question 3.** Does the descriptor, *never*, establish measurement equivalence on the initial threshold of the scale of the construct of interest for the screener across each of the racial and ethnic groups relative to the European American index group?

The scalar model nests within the metric model by constraining both the thresholds and the factor loadings. The multigroup scalar model on both the ESBA and the SRSS maintained the same as the metric model CFI of 0.967 and 0.971, respectively. The multigroup scalar model of the SIBS improved from the metric model by .001. The ESBA, the SRSS and the SIBS held the same origin on the scale, meaning that the ordinal descriptor *never* held the same starting level of frequency across the racial and ethnic
groups with the ESBA ($\chi^2$ (250) 449.129, $p < .001$, RMSEA = 0.032, CFI = 0.928), the SRSS ($\chi^2$ (90) 190.861, $p < .001$, RMSEA = 0.038, CFI = 0.971), and the SIBS ($\chi^2$ (90) 136.743, $p = .001$, RMSEA = 0.026, CFI = 0.979). The screeners demonstrated scalar invariance.

**Details of the analysis.** The results from the multigroup confirmatory factor analysis indicated that each screener satisfied tests for measurement invariance. Further analysis of the research questions highlighted the strengths and strains on each model for each screener at successive levels of testing. The following analysis examines each screener for each research question.

**Research question 1.1.** The initial unconstrained Elementary Social Behavior Assessment (ESBA), which measures level of teachability, demonstrated a poor fit across the chi square test of exact fit, the RMSEA statistic for close fit, and the CFI for incremental fit. Four residuals in the initial omnibus ESBA model exhibited positive polychoric correlations above .10, suggesting the need to respecify the model with covarying residuals (Appendix G). The modification indices for the first respecified ESBA model indicated significant improvement in the fit with covarying another eight residuals. The 12 covarying residuals make sense conceptually because they align with the underlying theory of academic enablers. Such items may be easily confused as measuring the same concept, and as a result the covariance not explained by the latent variable of teachability may arise from the inability by the teacher to distinguish the items. The final respecified omnibus ESBA model reached a marginal fit with $\chi^2$ (42) 461.184, $p < .001$; RMSEA = 0.050; CFI = 0.931. The RMSEA (90%CI = 0.046 – 0.055) suggested a close fit along with the CFI within an acceptable fit.
Table 10
*Tests of Measurement Invariance on Respecified ESBA with Covarying Residuals*

<table>
<thead>
<tr>
<th>Measurement Invariance</th>
<th>$\chi^2$ (df)</th>
<th>RMSEA (90% CI)</th>
<th>$p_{RMSEA \leq .05}$</th>
<th>$\Delta$ in CFI</th>
<th>Same =&lt; .01</th>
<th>diff &gt;.01</th>
<th>Different</th>
</tr>
</thead>
<tbody>
<tr>
<td>Configural</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(unconstrained loadings and thresholds)</td>
<td>651.080 (210)</td>
<td>0.052</td>
<td>(0.047, 0.056)</td>
<td>0.928</td>
<td>na</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metric</td>
<td>454.715 (254)</td>
<td>0.032</td>
<td>(0.027, 0.036)</td>
<td>0.967</td>
<td>-.039</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Scalar</td>
<td>449.129 (250)</td>
<td>0.032</td>
<td>(0.027, 0.037)</td>
<td>0.967</td>
<td>0.0</td>
<td>No</td>
<td></td>
</tr>
</tbody>
</table>

The respecified ESBA model appears to measure the latent construct of teachability with the equivalent factor structure across each of the racial and ethnic groups relative to the European American index group. Table 10 shows the test of configural invariance for the respecified ESBA model maintained a marginal fit with $\chi^2$ (210) 651.080, $p < .001$; RMSEA = 0.052; CFI = 0.928. The standardized factor loadings exceeded 0.680 for all indicators across each racial and ethnic group.

**Research question 1.2.** The Student Risk Screening Scale (SRSS) measures the construct of externalizing problems with the equivalent factor structure for each of the
racial and ethnic groups. The test for configural equivalence on the SRSS attained a good fit ($\chi^2 (70) = 256.245, p < .001, \text{RMSEA} = 0.058, \text{CFI} = 0.947$; Table 11).

Table 11
Tests of Measurement Invariance on Student Risk Screening Scale

<table>
<thead>
<tr>
<th>Measurement Invariance</th>
<th>$\chi^2$ (df)</th>
<th>RMSEA (90% CI)</th>
<th>$p$ RMSEA $\leq .05$</th>
<th>CFI</th>
<th>$\Delta$ in CFI</th>
<th>Same $= &lt; .01$</th>
<th>diff $&gt;.01$</th>
<th>Different</th>
</tr>
</thead>
<tbody>
<tr>
<td>Configural (unconstrained thresholds and loadings)</td>
<td>256.245 (70)</td>
<td>0.058 (0.051, 0.066)</td>
<td>0.947</td>
<td>na</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metric (constrained loadings)</td>
<td>197.078 (94)</td>
<td>0.038 (0.030, 0.045)</td>
<td>0.971</td>
<td>-0.024</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scalar (constrained thresholds and loadings)</td>
<td>190.861 (90)</td>
<td>0.038 (0.030, 0.045)</td>
<td>0.971</td>
<td>0.0</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Items E1 and E5 presented the lowest standardized loadings. Specifically, the ratings on the frequency of academic problems (E5) obtained standardized loadings between a low of .346 for Latino Hispanic students and a high of .482 for multiracial students. The next lowest item on standardized loadings (.526 – .565) was stealing (E1). Otherwise, the remaining items demonstrated loadings above .645.

**Research Question 1.3.** The Student Internalizing Behavior Scale (SIBS) measures the construct of internalizing problems with the equivalent factor structure across each of the groups. The test for configural equivalence on the SIBS attained a good fit ($\chi^2 (70) = 190.867, p < .001, \text{RMSEA} = 0.047, \text{CFI} = 0.945$; Table 12).
Table 12
Tests of Measurement Invariance on the Student Internalizing Behavior Scale

<table>
<thead>
<tr>
<th>Measurement Invariance</th>
<th>$\chi^2$ (df)</th>
<th>RMSEA (90% CI)</th>
<th>$p_{\text{RMSEA} \leq .05}$</th>
<th>$\Delta$ in CFI</th>
<th>Same $\leq .01$</th>
<th>Different</th>
</tr>
</thead>
<tbody>
<tr>
<td>Configural</td>
<td>190.867 (70)</td>
<td>0.047 (0.039, 0.055)</td>
<td>$p = 0.729$</td>
<td>0.945</td>
<td>na</td>
<td></td>
</tr>
<tr>
<td>Metric</td>
<td>141.770 (94)</td>
<td>0.025 (0.016, 0.034)</td>
<td>$p = 1.000$</td>
<td>0.978</td>
<td>-0.033</td>
<td>No</td>
</tr>
<tr>
<td>Scalar</td>
<td>136.743 (90)</td>
<td>0.026 (0.016, 0.034)</td>
<td>$p = 1.000$</td>
<td>0.979</td>
<td>-0.001</td>
<td>No</td>
</tr>
</tbody>
</table>

The range of standardized loadings on the SIBS for low academic achievement appeared the lowest at .365 for Asian American and the highest of .464 for both European American and multiracial students. These standardized factor loadings on the SIBS differed from those on the SRSS on the same item content. The next lowest standardized loading, complains about being sick or hurt, had a range of .433 to .577. The remaining items exhibited loadings above .620.

Research question 2: Metric invariance. Do the ordinal descriptors of never, occasionally, sometimes, and frequently demonstrate measurement equivalence in the metric unit of frequency on for latent construct of interest for the screener across each of the racial and ethnic groups relative to the European American index group?

To address this question, metric invariance involves two steps. The first step begins by setting an equality constraint on each indicator across each of the racial and
ethnic groups. The second step compares the change in the CFI determines the degree of difference between the unconstrained configural model and the constrained metric model. If the change in CFI between the two models is trivial (<.01;), then the more constrained model is retained (Cheung & Rensvold, 2002). As the constrained model becomes more complex, usually the fit among models becomes harder to attain (Kang et al., 2016).

Surprisingly, the model for each screener improved by constraining the factor loadings. The unexpected improvement in the goodness of fit during multigroup tests of measurement invariance with ordinal ratings has occurred in other studies (Lai & Yoon, 2015; Pendergast, von der Embse, Kilgus, & Eklund, 2017). Each screener demonstrates metric invariance, meaning the ordinal descriptors of never, occasionally, sometimes, and frequently share the same meaning across each group with factor loadings from European American group serving as the index.

**Research question 2.1.** On the ESBA, the difference between the configural CFI of 0.928 and the metric CFI of 0.967 was -.039 (Table 10). The lowest standardized factor loading across all of the indicators was works with effort (.661) for Asian American students. The ESBA demonstrates metric equivalence for the ordinal descriptors for latent construct of teachability with \( \chi^2 (254) = 454.715, p < .001 \), RMSEA = 0.032, CFI = 0.967.

**Research question 2.2.** On the SRSS, the difference between the configural CFI of 0.971 and the metric CFI of 0.967 was -0.024 (Table 11). With the equality constraints on the factor loadings of the SRSS, the standardized loadings for low academic achievement appeared weakest at .293 for Asian American students and highest for multiracial students at .466. Moderate loadings appeared in stealing and peer
Despite the strain in the model from E5, the SRSS demonstrated metric equivalence for the ordinal descriptors for latent construct of externalizing problems with $\chi^2 (94) = 197.078$, $p < .001$, RMSEA = 0.038, CFI = 0.971.

**Research question 2.3.** On the SIBS, the difference between the configural CFI of 0.945 and the metric CFI of 0.978 was -0.033 (Table 12). The greatest strain in the metric model occurred with *low academic achievement* and *complains about being sick or hurt*. The standardized factor loading on *low academic achievement* on the SIBS continued to rate higher than on the SRSS, with a range from the lowest for Asian American students at .364 to the highest for European American students at .475. The standardized loading for *complains about being sick or hurt* went from .460 for Latino Hispanic students to .513 for Asian American/Pacific Islanders. Items I1, I2, and I3 show moderate loadings. The SIBS demonstrated metric equivalence for the ordinal descriptors for latent construct of internalizing problems with $\chi^2 (94) = 141.770$, $p = .001$, RMSEA = 0.025, CFI = 0.978.

**Research question 3: Scalar invariance.** Scalar invariance involves constraining both the thresholds and the factor loadings across groups. Tests for scalar invariance for ordinal data use thresholds (rather than intercepts). When the thresholds between *never*, *occasionally*, *sometimes*, and *frequently* are similar across groups, we can say that differences from the scale score arise from “true” differences in the latent construct. The change in CFI based on the difference between the scalar model and the metric model determines whether the fit between the equality-constrained loadings and thresholds model degrades from the model with equality-constrained loadings only.
**Research question 3.1.** On the ESBA, the scalar CFI remained 0.967 despite the loss of four degrees of freedom between the metric and scalar models (Table 10). The change in the upper limit of the confidence interval for the RMSEA by 0.001 reflects the change in model fit due to the loss of the four degrees of freedom. The largest differences in constrained thresholds occurred between African American and Asian American/Pacific Islander on *works out strong feelings appropriately* (.411) and *resolves peer conflict* (.432). The two modification indexes above 3.84 did not make sense conceptually, suggesting that the test for scalar invariance did not present any remarkable strains in the ESBA model. The ESBA demonstrated scalar equivalence with the descriptors, *never, occasionally, sometimes*, and *frequently*, on the initial threshold of the scale of teachability with $\chi^2 (250) = 449.129, p < .001$, RMSEA = 0.032, CFI = 0.967. Appendix H presents the tables of the final parameter estimates for scalar invariance and the scalar model of the ESBA for each racial and ethnic group.

**Research question 3.2.** On the SRSS, the scalar CFI stayed at 0.971 despite the loss of four degrees of freedom between the metric and scalar models (Table 11). The change in the $p$ value for the RMSEA < .05 indicated that minor change in the complexity of the scalar model produced a very small adjustment. The largest differences in constrained thresholds on the SRSS appeared between African American and Asian American/Pacific Islander students for *lie, cheat, sneak* (.428), *behavior problems* (.640), *low academic achievement* (.400), and *negative attitude* (.520). The standardized threshold on *stealing* for Asian American students appeared elevated compared to that for the other racial and ethnic groups. This increase in the standardized threshold occurred because none of the Asian American/Pacific Islander students
received a rating of *frequently* on the SRSS, leading to an empty cell (zero cell) for the four ordinal categories. In effect, the standardized intercept increased for Asian American/pacific Islander students due to a scale of three ordinal categories (i.e., *never*, *occasionally*, and *sometimes*). Despite the zero cell in the frequency of the *frequent* rating in the scalar model, the SRSS demonstrated scalar equivalence with the descriptors, *never, occasionally, sometimes, and frequently*, on the initial threshold of the scale of externalizing problems with $\chi^2 (90) = 190.861$, $p < .001$, RMSEA = 0.038, CFI = 0.971. Appendix I presents the tables of the final parameter estimates for scalar invariance and the scalar model of the SRSS for each racial and ethnic group.

**Research question 3.3.** On the SIBS, the difference between the metric CFI of 0.978 and the scalar CFI of 0.979 was -0.001 (Table 12). The RMSEA also improved by 0.001 due to the increased parsimony from the loss of four degrees of freedom. The constrained thresholds on *low academic achievement* exhibited a difference of .434 between African American and Asian American/Pacific Islander students. The ESBA demonstrated scalar equivalence for the descriptor, *never*, on the initial threshold of the scale of internalizing problems with $\chi^2 (90) = 136.743$, $p = .001$, RMSEA = 0.026, CFI = 0.979. Appendix J presents the tables of the final parameter estimates for scalar invariance and the scalar model of the SIBS for each racial and ethnic group.

The fit from the metric model to the scalar model either remained the same or improved. None of the parameters for the scalar models exhibited presented strain. Therefore, the meaning of the ordinal descriptors, *never, occasionally, sometimes, and frequently*, held the same meaning across groups.
Summary

This study analyzed the degree of measurement invariance for three mental health screeners across five racial and ethnic groups. The ESBA measures level of teachability. The SRSS quantifies the level of risk for externalizing problems. The SIBS measures the level of risk for internalizing problems. The data used in this analysis come from teacher ratings of 3965 elementary school students.

The initial unconstrained baseline models demonstrated adequate fit for each separate group of African American, Asian American, Latino Hispanic, multiracial, and European American students. The tests for measurement invariance simultaneously compared each racial and ethnic group to the European American index group. Each step in testing for measurement invariance increased the complexity of the model by adding constraints. The first test for configural invariance tests established the unconstrained foundation for comparing the subsequent test of metric invariance in which the factor loadings were constrained. The models with metric invariance compared to the models with both thresholds and factor loadings were constrained to test for scalar invariance.

On the first research question, the ESBA, the SRSS, and the SIBS demonstrated configural invariance. Each screener measures the same latent construct across the five racial and ethnic groups because they share the same constellation of meaningful factor loadings. For the second research question, the ESBA, the SRSS, and the SIBS demonstrated metric invariance. Each racial and ethnic group shared similarities in their factor loadings on each screener compared to the European American reference group. Metric equivalence indicates that the slope for each unit of change from never to occasionally to sometimes to frequently holds the same metric unit across the five groups.
On the third research question, the ESBA, the SRSS, and the SIBS demonstrated scalar invariance. Each racial and ethnic group shared equivalent thresholds and factor loadings compared to the European American index group. Scalar equivalence indicates that the origin for never, occasionally, sometimes, and frequently start at the same point for each of the five groups.
Chapter Five: Discussion

The population of elementary schools in the United States is becoming more racially and ethnically diverse. Simultaneously and positively, schools are increasing support for students with mental health needs. Mental health screening assists decision-making by providing a reliable and valid means of identifying students in need of mental health supports. The fairness of early identification of mental health needs across diverse racial and ethnic groups depends, in part, on the psychometric property of measurement invariance. Specifically, the mental health screening instruments need to demonstrate that they share the equivalent meaning for the ordinal ratings of never, occasionally, sometimes, and frequently and the equivalent origin of scale across each racial and ethnic group.

The purpose of this study was to test for measurement invariance simultaneously across African American, Asian American, Latino Hispanic, and multiracial students compared with European American students on Elementary Social Behavior Assessment (ESBA; Pennefather & Smolkowski, 2015), the Student Risk Screening Scale (SRSS; Drummond, 1994), and the Student Internalizing Behavior Screener (SIBS; Cook et al., 2011).

This chapter discusses the conclusions drawn from the multigroup confirmatory factor analyses for measurement invariance. The results have important implications for students, teachers, parents, and policymakers. Moreover, the results have important implications related to disproportionality in special education and school discipline. Despite the limitations of this study and the need for additional research, the results of
this study suggest that the ESBA, SRSS, and SIBS do not exhibit biased measurement depending on the race and ethnicity of the student.

Summary of the Results

The size of the sample adequately satisfied requirements for sufficient power to detect differences in the models for each racial and ethnic group. The missing data affected only 2% of total items. The patterns of missing data reflected classrooms that did not complete entire selected screeners. The data demonstrate severe skewness and kurtosis. Compared to the national composition of elementary schools, the sample included more than double Asian American/Pacific Islanders, four times multiracial students, along with approximately similar levels of African American and Latino Hispanic students. The distributions of students across grade levels and between genders are nearly equivalent. The results from this study may generalize across culturally diverse elementary school students. Nonetheless, the ESBA needs replication with another sample to prevent capitalizing on chance as a result of the respecification.

The research questions tested whether the ESBA, SRSS, and SIBS measured equivalent constructs with equivalent measurement properties. Multigroup confirmatory factor analysis simultaneously compared African American, Asian American, Latino Hispanic, and multiracial groups of students with the European American group. The first research question tested whether the ESBA, SRSS, and SIBS measure their respective constructs with equivalent factor structure with meaningful (nontrivial) factor loadings for each racial and ethnic group. The ESBA required respecification to establish an adequate baseline model for the construct of teachability. Each screener demonstrated
configural invariance in which the indicators for each latent construct function in a similar manner for each racial and ethnic group in the sample.

Using the configural model as the starting point for comparisons, the second research question tested whether the ordinal ratings of the indicators on the ESBA, SRSS, and SIBS measure the equivalent meaning for each metric unit. For each screener, the results indicated that the ordinal descriptors of *never, occasionally, sometimes,* and *frequently* held equivalent values in each racial and ethnic group. The third research question tested whether the origin for the scale of measurement starts with the equivalent value on the ESBA, SRSS, and SIBS. The results established scalar invariance in which the ordinal descriptors for each screener start with the same initial value of the scale.

Previously none of these screeners had empirically demonstrated measurement invariance, and this study fills the gap by establishing scalar invariance for the ESBA, SRSS, and SIBS. Teachers and researchers can have confidence that these screeners share the equivalent meaning in the construct and equivalent psychometric properties across African American, Asian American, Latino Hispanic, European American and multiracial groups of students.

**Discussion of the Results**

Measurement invariance provides evidence of reliability and validity of the scores across the sampled racial and ethnic groups (Chen, 2008). Measurement invariance reduces some bias in the measurement of the latent construct. Less bias improves the validity of the interpretation of the scores. Measurement invariance enables generalization across represented populations. Screeners with measurement invariance
help teachers identify students from diverse backgrounds who are healthy and students who truly need additional support.

Examining for measurement invariance uses multigroup confirmatory factor analysis to compare a series of nested models in which equality constraints on parameters are added onto successive tests. The instruments demonstrate invariance when membership in a racial or ethnic group does not influence the parameter estimates in each of the research questions. In other words, the relationship between the latent variable to its indicators demonstrates equivalency between racial and ethnic groups.

**Research question 1.** The results from first research question established configural invariance for ESBA, SRSS, and SIBS, indicating that the screeners measure their construct with same meaning for each racial and ethnic group (Hui & Trandis, 1985). Configural invariance indicates the degree of overlap in measuring the construct in each racial and ethnic group. For instance, the lack of adequate overlap due to construct-irrelevant factors increases bias. In comparison, the degree of overlap on construct-relevant behaviors for each racial and ethnic group establishes the pattern of configural invariance. In practice, configural invariance means that the screeners compare apples to apples and not apples to oranges (Stegmueller, 2011).

Configural invariance on the ESBA means that the construct of teachability influences interpersonal skills, study skills, motivation, and engagement in a similar manner when applied to racial and ethnic groups. The strong standardized factor loadings on the items indicate the effect of teachability. However, some of the manifestations of teachability appear difficult to distinguish from one another, such as a student *working with effort* entails *doing the seatwork as directed*. At least ten pairs of
items on the ESBA share such highly correlated residuals in measurement of the construct of teachability.

Configural invariance on the SRSS indicates the construct of externalizing problems captures the meaning of antisocial and disrespectful classes of behavior (Gresham, 2015) with moderate to strong standardized factor loadings for each of the sampled groups. However, low academic achievement presented a weak association with externalizing problems among the seven indicators. Thus, students may not appear to struggle academically to the same degree as with externalizing problems.

Configural invariance on the SIBS indicates that the construct of internalizing problems associated with anxiety and depression maintains moderate to strong factor loadings across each of the racial and ethnic groups. Just like in the SRSS, low academic achievement presented a weak association with internalizing problems, suggesting that internalizing problems may not contribute to low academic achievement. The initial models of the ESBA, SRSS, and SIBS with configural invariance establish adequate starting points from which to compare more constrained models in the second research question.

**Research question 2.** The ESBA, the SRSS, and the SIBS demonstrate metric invariance. Testing for metric invariance involves setting the factor loadings to equality constraints with equal values for the slopes. Factor loadings function like regression slopes between the indicators and latent constructs, presenting the magnitude of change in the indicator for each unit of change in the factor. Metric invariance means that each unit of change in the indicator corresponds to a consistent change in the level of construct of interest without the student’s racial or ethnic background influencing the
interpretation of *never*, *occasionally*, *sometimes*, and *frequently*. The results of the equal factor loadings provide evidence that the model reproduces the observed relationships among the indicators across the racial and ethnic groups. Furthermore, the results from the second research question established metric invariance for the ESBA, SRSS, and SIBS in which the meaning of the metric unit in *never*, *occasionally*, *sometimes*, and *frequently* holds equivalent values for each racial and ethnic group.

Metric invariance means that the unit of measurement is identical for each racial and ethnic group (Vandenberg & Lance, 2000). Each of the screeners uses ordinal descriptors of *never*, *occasionally*, *sometimes*, and *frequently* to measure the frequency of the occurrence of the indicators. Ordinal descriptors do not fix the magnitude for each level of frequency. These ordinal descriptors imply that *sometimes* occurs more often than *occasionally*. Thus, the looseness in the magnitude for ordinal descriptors introduces potential bias into the measurement of the indicators. In practice, teachers interpret the ordinal descriptors in the same manner applied to each racial and ethnic group. Metric invariance establishes the minimum level of construct validity for interpreting the screeners in an equivalent manner across each racial and ethnic group.

The factor loading on *low academic achievement* demonstrated the weakest association of externalizing problems and internalizing problems. Specifically, the standardized factor loading on *low academic achievement* dropped to .293 on externalizing problems for Asian American students compared to the highest loading for multiracial racial students at .466. Similarly, the standardized factor loading on *low academic achievement* reached .364 on internalizing problems for Asian American students compared to the highest loading for European American students at .460.
Basically, externalizing or internalizing problems associate weakly with low academic achievement across the groups.

**Research question 3.** The ESBA, the SRSS, and the SIBS demonstrate scalar invariance. Testing for scalar invariance involves setting the threshold and the factor loadings to equality constraints. Thresholds function like a regression intercept as the starting value between the indicator and the latent variable. Constraining the thresholds means that the underlying intercepts for the unstandardized loadings are set to equal values. Scalar invariance permits the interpretation of the total score on each screener to hold the same meaning for each racial and ethnic group. In effect, the total score with scalar invariance does not introduce bias into the determination of the cut-score for risk status. This scalar invariance allows unbiased comparisons of total scores among African American, Asian American, Latino Hispanic, and multiracial students compared with European American students. Without scalar invariance, differences in total scores could arise from differences in the thresholds for the indicators. In the absence of scalar invariance, the total score used for a cut-off for selection for intervention would be more accurate for some groups and less accurate for other groups, leading to misclassification (Millsap & Kwok, 2004). Fair and equitable use of total scores requires scalar invariance across racial and ethnic groups (Meredith, 1993; Meredith & Teresi, 2006).

Although the screeners demonstrate scalar invariance, differences in the constrained thresholds occur. Specifically, the most pronounced differences occur between African American and Asian American groups of students. On the ESBA, the largest differences among constrained thresholds appear on *works out strong feelings appropriately* and *resolves peer conflict*. On the SRSS, the largest differences occur on
behavior problems and negative attitude. On the SIBS, the largest difference appears on low academic achievement. On face value, the differences on the thresholds for these specific items may appear to matter. However, the magnitude of these differences does not affect the underlying true score of African American and Asian American groups of students (Millsap & Kwok, 2004).

The SRSS measures antisocial behaviors violate social norms. The covert behaviors such as stealing appear more difficult to rate via frequency compared to the overt ones such as fighting. Specifically, none of the 539 Asian American students received a rating of frequently for stealing, although this zero level in the teacher rating of covert behavior suggests a cultural difference in covert violation of social norms. Nonetheless, such a cultural artifact does not make a difference in the measurement of the construct when simultaneously comparing the diverse groups.

The results of this study add to the evidence of construct validity for the ESBA, SRSS, and SIBS. Screening instruments with scalar invariance across diverse populations demonstrate construct validity through equivalent unit and equivalent origin for the ordinal descriptors (Millsap, 2011). The ESBA, SRSS, and SIBS reflect the same construct of interest and the total score indicates the same meaning across the racial and ethnic groups (AERA et al., 2014). Differences in total scores on the screeners across racial and ethnic groups may arise from differences in the true levels of the latent construct (Millsap & Olivera-Aguilar, 2012) and not from artifacts of measurement (Meredith & Teresi, 2006). The purpose of measurement invariance is to assist in making valid comparisons across diverse racial and ethnic groups (Bowen & Masa, 2015;
Meredith, 1993; Sass, 2011). The total scores from ESBA, SRSS, and SIBS generalize across racial and ethnic groups and reduce bias in the referrals for assistance.

The reduced biased in the scores due to measurement invariance on the ESBA, SRSS, and SIBS enables better precision and accuracy to identify students in need of social, emotional, or behavioral supports (Millsap & Kwok, 2004). Least biased assessments are less likely to over-estimate or under-estimate the level of the construct across diverse cultures (Millsap, 2007). “Unless measurement invariance holds, fairness and equity cannot exist in principle. Thus, when the purpose of the test use is selection of individuals, measurement invariance is a necessary condition for fair selection procedures” (Boorsboom, 2006, p.179). Thus, in the case of screening diverse racial-ethnic groups for selection for intervention, measurement invariance reduces the likelihood that healthy kids are falsely identified and that kids in need of support fall through the cracks (Chen, 2008). Establishing measurement invariance on the ESBA, SRSS, and SIBS leads to important opportunities for students and teachers.

Implications at the student and teacher level. The implications for screeners with measurement invariance potentially influence a cascade of effects to improve the outcomes of students and teachers. Teachers can make valid decisions from them. Teachers can confidently use the ESBA, SRSS, and SIBS to identify racially and ethnically diverse students in need of social, emotional, and behavioral supports. Importantly, teachers can use the information from the screeners to provide culturally responsive prevention and intervention. Students may benefit from more equitable opportunities to develop social, emotional, and behavioral health as a result of educating the whole child. Ultimately, by addressing social, emotional, and behavioral needs
through positive development, teachers may reduce disproportionality in referrals to discipline and special education.

Universal screeners improve the detection of minor levels of problems that often occur within a teacher’s threshold for tolerance for teachability (Weist et al., 2007). For example, the level of teachability on the ESBA reflects the student’s standing in relation to the teacher’s threshold for tolerance for behavioral diversity. Teachers can use the ESBA to identify students struggling with interpersonal skills, study skills, motivation, and engagement in learning. Teachers may prevent students from failing by providing direct instruction in academic enabling skills that satisfy their expectations for teachability (Marquez et al., 2014). This demystification of teacher’s expectations enables equitable access to participation in and benefit from learning within the classroom culture of power (Delpit, 1988).

Universal mental health screening increases the likelihood of early identification of social, emotional, and behavioral needs before they blossom into mental disorders (Brown & Barlow, 2005; Dowdy et al., 2015). The SRSS and SIBS help teachers identify students whose levels of externalizing and internalizing problems appear at subclinical levels and, therefore, may not exceed the threshold for teachability. For example, the teacher could use information from the SRSS to proactively adjust classroom management practices. Specifically, the teacher could arrange seating assignments that promote positive peer reporting (Babyak, Luze, Kamps 2000; Bowman-Perrott, Burke, Zaini, Zhang, & Vannest, 2016) with culturally responsive group contingencies from the Good Behavior Game (Nolan, Houlihan, Wanzek, & Jenson, 2014). In such instances, the SRSS helped the teacher intervene early rather than wait for the student to fail. Early
prevention addressing externalizing problems allows teachers to proactively implement culturally responsive instructional practices rather than react with disciplinary practices that could jeopardize healthy student-teacher relationships (Klingner et al., 2005; Saifer, Edwards, Ellis, Ko, & Stuczynski, 2011).

Similarly, the SIBS helps teachers to identify students with internalizing problems who perform within the threshold of teachability. Teachers often overlook students with internalizing problems because they do not interrupt the flow of instruction to the same degree as students with externalizing problems (Gresham & Kern, 2004; Kahlberg, Lane, Driscoll, & Wehby, 2011; Liljequist & Renk, 2007; Skiba, Michael, Nardo, & Peterson, 2002). Yet, the SIBS focuses teacher attention toward observable behaviors associated with multicultural research regarding cultural experiences of internalizing disorders (Anderson & Mayes, 2010; Austin, & Chorpita, 2004; Ritsher, Struening, Hellman, & Guardino, 2002). For example, African American, Asian American, and Latino Hispanic youth report somatic complaints as culturally appropriate expression of internalizing symptoms (Banh et al., 2012; Choi & Park, 2006). The teacher can use the information from the SIBS to promote mental health practices such as social emotional learning in their classroom (Franklin et al., 2012). In addition, teachers integrate social and emotional learning skills into classroom routines that cultivate self-awareness, self-management, social awareness, relationship skills, and responsible decision-making (Durlak et al., 2011; Greenberg et al., 2003; Taylor, Oberle, Durlak, & Weissberg, 2017). This proactive approach may improve equity and fairness by addressing mental health before students surpass the teacher’s threshold for tolerance, resulting in a referral for discipline or special education (Raines et al., 2012).
Early identification opens opportunities to intervene early and prevent the development of chronic and severe mental health problems while promoting equal opportunity to develop social, emotional, and behavioral health (Cruden et al., 2016; Greenberg et al., 2001). For racially and ethnically diverse students with unmet mental health needs, early identification of social, emotional, and behavioral needs enables students to access intervention services. Using the combination of information from the ESBA, SRSS, and SIBS, teachers can tailor interventions to the specific strengths and challenges of students based on their levels of teachability and internalizing and/or externalizing problems (O’Connell et al., 2009; Walker & Gresham, 2015). The ESBA, SRSS, and SIBS provide teachers with more information about the whole child so that considerations for interventions become more likely at vulnerable decision points compared to waiting for the student to fail, resulting in referrals for school discipline or special education services (Smolkowski, Girvan, McIntosh, Nese, & Horner, 2016).

**Implications at the school and policy level.** Schools and policy makers can use the data from the ESBA, SRSS, and SIBS to identify systemic needs of students and to measure the effects of prevention and intervention initiatives (Dowdy et al., 2015). Schools using the ESBA, SRSS, and SIBS can target interventions to specific risk factors and evaluate the effects of the interventions while taking racial and ethnic diversity into account (Lyon, Maras, Pate, Igusa, & Vander Stoep, 2016). For example, the aggregated results from least biased screeners may indicate disparate systemic opportunities to learn if underidentification and overidentification of diverse students continue to occur despite targeted interventions (VanDerHeyden & Witt, 2005). Aggregate data from least biased universal mental health screeners allow schools to identify climate conditions that expand
or depress the opportunity to develop social, emotional, and behavioral well-being (Skiba, Knesting, & Bush, 2002).

Universal screening supports a public health approach to addressing unmet mental health needs (Stiffman et al., 2010). Schools increasingly serve as the primary provider of mental health services for children (Bruhn, Woods-Groves, & Huddle, 2014; Merikangas et al., 2011). Mental health services integrate into multi-tiered systems of support (MTSS). Schoolwide positive behavioral supports link community and school mental health services in the integrated systems framework (Barrett, Eber, & Weist, 2013). Similarly, the comprehensive, integrated three-tiered model creates a unified system to address academic, behavioral, and social supports (Lane, Oakes, Ennis, & Hirsch, 2014). School-based mental health services have contributed to a reduction in disproportionality in school discipline (Bruns, Moore, Stephan, Pruitt, & Weist, 2005), especially for African American youth (Darensbourg et al., 2010). School-based mental health services contributed to improved academic outcomes (Walker, Kerns, Lyon, Bruns, & Cosgrove, 2010).

The ESBA, SRSS, and SIBS play a role in the culturally competent assessment system (Knight & Zerr, 2010). Screening instruments with measurement invariance contribute to fair and equitable assessment for students from diverse racial and ethnic backgrounds (Flanagan, Ortiz, & Alfonso, 2013). By reducing bias in the instrument, teachers and administrators can focus on culturally competent interpretation of the results that measure their efforts toward educational equity at the systems level (Skiba et al., 2002). Using least biased screeners permits reframing the problem of disproportionality through a cultural-historical lens that allows educators to critically analyze the role of
cultural practices in decision-making within the historical context of the opportunity to learn (Artiles, 2009; Lee, 2009). Specifically, early identification of needs presents the opportunity to foster resiliency and skills that prevent the development of deficits. The ESBA, SRSS, and SIBS demonstrate cross-cultural equivalency that can promote culturally responsive pedagogy by reducing the placement of “misidentified, misplaced, and misunderstood” culturally diverse students into special education (Harris-Murri, King, & Rostenberg, 2006).

**Limitations**

Limitations to this study involve cultural, methodological, and ethical issues. Two cultural concerns include the use of pan-ethnic groups for generalization and the use of European American students for the reference group. The methodological issues include the constraints of statistical analysis of ordinal data and the use of secondary data. The ethical issues acknowledge the concern for community acceptance of universal mental screeners and the right to privacy.

**Cultural considerations.** This study conceptualized racial and ethnic groups based on identification within the categories provided by the school district. These pan-ethnic models of racial and ethnic categories may misrepresent African Americans, Asian Americans, and Latino Hispanics as homogeneous groups (Marks, Ejesi, & Garcia Coll, 2014). Pan-ethnic models often do not reflect within-group heterogeneity involving history, culture, and immigration experiences. The pan-ethnic African American group includes Caribbean-born and immigrants from Africa. The pan-ethnic Asian American group includes diverse people from southeast Asia (e.g., Cambodia and Viet Nam) to east Asia (e.g., Taiwan, China, Korea, & Japan), and to the Pacific Islands (e.g., the
Philippines). The Latino Hispanic group includes people from Puerto Rico and the Caribbean Islands to South America, Central America, and Mexico. The within-group heterogeneity among the pan-ethnic groups in this study masks the potential differences in teachability, externalizing and internalizing problems. Analyzing pan-ethnic data underestimates the heterogeneity within each racial and ethnic group (Artiles, Rueda, Salazar, & Higareda, 2005).

Cultural concepts of distress vary from group to group (Office of the Surgeon General, 2001). The within-group heterogeneity among the pan-ethnic groups in this study masks the potential differences in teachability, externalizing and internalizing problems. Analyzing pan-ethnic data underestimates the heterogeneity within each racial and ethnic group (Artiles, Rueda, Salazar, & Higareda, 2005). As a result of cultural differences in the experience and expression of distress, mental health problems may persist unrecognized (Li & Seidman, 2010; Lindsey, Brown, & Cunningham, 2017; Toppelberg, Hollinshead, Collins, & Nieto-Castañon, 2013).

This study used a cross-cultural design with the European American students as the reference group for comparison among African American, Asian American, Latino Hispanic, and multiracial students. According to the Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service, 82% of teachers in the United States identify as European American (U.S. Department of Education, 2016a). Thus, the majority of teachers may hold expectations of European American cultural norms. However, using European American students as the reference group may lead teachers to continue to see cultural differences through the lens of deficits.
**Methodological considerations.** This study used a two-stage sampling method that did not adjust for non-independence of clustered sampling. In the first stage, schools met the criteria for selection. In the second stage, within each school, teachers rated their classrooms of students. The statistical applications for *lavaan* (Rosseel et al., 2017), *lavaan.survey* (Oberski, 2014, 2016) and *survey data* (Lumley, 2010, 2017) do not allow for survey weights of clustered samples using ordinal data.

The classroom clusters of students in the second stage of sampling are dependent upon their teacher’s rating (Stapleton, Yang, & Hancock, 2016). The extant data do not address classroom-level membership, precluding the calculation of the intra-cluster correlation (ICC). Nonetheless, the level of fidelity of implementation of Schoolwide Positive Behavior Interventions and Supports indicates that the selected schools may appear more similar, possibly contributing to reduced variance between classroom clusters. Additionally, the mixture of five racial and ethnic groups in each classroom may increase the within-cluster variance, thereby reducing the effects of the intra-cluster correlation within the 161 classroom clusters with an average of 24 students (Laï & Kwok, 2015). Thus, the number of clusters in addition to the heterogeneity within clusters allows for the tolerance for effects of clustering (Maas & Hox, 2005).

Furthermore, the results from a simulation study indicate robustness of multigroup confirmatory factor analysis of measurement invariance with regard to dependent data (Perdomo et al., 2014).

The study used secondary data that had been collected prior to the initiation of this study. The director of implementation from the University of Washington (UW) described the data collection process. Without formal checks for the fidelity of the
administration of the screeners and the collection of the data, the adherence to the protocol and quality of the data collection remain uncertain. Nonetheless, this study examined only extant data from schools that met the inclusion criteria. Most of the administrators and teacher experienced fewer than three years of experience with school-wide positive behavior interventions and supports. This limited experience may have affected school climate in ways that increased or suppressed teachers responding consistently to behavior.

**Ethical considerations.** Ethical considerations in universal mental health screening include community acceptance, family privacy rights, and disproportionality in the identification of students in need of social, emotional, and behavioral supports (Chafouleas, Kilgus, & Wallach, 2010). The participating school district addresses these ethical concerns with universal mental health screening through their strategic plan for 2015 – 2020. The school district promoted a Whole Child initiative through a public awareness campaign, addressing well-being, including mental health, as an essential element of learning. The school district’s Office of Community Partnerships involved neighborhood organizations in the campaign to represent a broad range of stakeholders. Furthermore, the district used the results from the Healthy Youth Survey from the secondary schools to justify universal mental health screening in elementary schools. Although the Healthy Youth Survey does not permit follow-up with students, the baseline from the surveillance data established the need to identify students at risk for mental health problems while they attended elementary school. The participating school district shared data from the pilot project on universal mental screening with community mental health service providers and the University of Washington – Tacoma, Whole Child
Initiative before going to scale with school selection. The community involvement in the Whole Child initiative along with the baseline data from the Healthy Youth Survey may have contributed to the acceptance of the need for universal mental health screening in elementary school.

Communities may oppose universal mental health screening due to the stigma associated with mental health problems (Carter, Briggs-Gowan, & Davis, 2004). Ethnically and racially diverse families may perceive school-based mental health services as less stigmatizing (Alegria, Vallas, & Pumariega, 2010). Records indicate that the procedures in the Whole Child initiative intended to preserve confidentiality while addressing concerns of stakeholders. For example, UW project administrators trained staff on the purpose of the screeners and family privacy rights in addition to addressing teachers’ concerns regarding the use of and access to the information. The school district addressed confidentiality of the screeners through the administration protocol. The district controlled access to the screening data to teacher leadership teams that used to data to ethically inform social, emotional, and behavioral interventions (Albers, Glover, & Kratochwill, 2007).

Parental consent for universal mental screening remains an area that requires attention (Blom-Hoffman et al., 2009; Chartier et al., 2008). The Protection of Pupil Rights Amendment of 1978 requires that parents provide written consent for surveys that involve sensitive information related to mental or psychological problems (20 U.S.C. § 1232h; 34 CFR Part 98). However, universal screening for instructional purposes, as part of the regular school activities does not require parental consent (IDEA; 2004; see § 34 C.F.R. 300.302 and § 34 C.F.R. 300.300[def][2][ii]). In the context of schools that have
demonstrated adequate levels of the implementation of School-wide Positive Behavioral Interventions and Supports, the data from universal screening procedures inform the instructional strategies used to address social, emotional, and behavioral development. Nonetheless, parental consent may be required for students to receive targeted or indicated interventions based on the information from the screening process. Therefore, school districts should create protocols for contacting parents regarding the purpose of the screeners and the protection of their privacy rights (Oakes et al., 2014). Schools should exert an effort to communicate the purpose of universal screening as the early identification of potential social, emotional, and behavioral challenges (Eklund & Kirgus, 2015).

**Recommendations for Further Research**

The aim of this study was to analyze universal mental health screeners for construct validity among diverse racial and ethnic groups. The ESBA, SRSS, and SIBS demonstrated scalar invariance across African American, Asian American, Latino Hispanic, and multiracial groups of students. Universal mental health screeners with measurement invariance support the research that addresses the increasing racial ethnic diversity of American school children. Such research intentionally addresses racial and ethnic variation so that prevention and intervention strategies may be adapted and developed for and delivered to diverse communities (Knight et al., 2009). These efforts focus on the effectiveness of culturally relevant research and practice.

Recommendations for further research include extending the study of construct validity to studies of predictive validity. This study emphasized the measurement properties of the latent constructs as a precondition for the study of the predictive value
of the constructs (Millsap, 2007). Establishing construct validity improves the interpretability of predictive validity (Betts et al., 2008). Screening instruments demonstrate predictive validity by analyzing the conditional probability of accurately identifying students who are truly at risk for poor outcomes (sensitivity) as well as students who are truly not at risk for poor outcomes (specificity). Using screeners with measurement invariance improves the diagnostic accuracy for predicting these outcomes across racial and ethnic groups.

Research studies on the predictive validity of the ESBA, SRSS, and SIBS may examine their effectiveness for accurately identifying racially and ethnically diverse students at risk. For example, just as the Social, Academic, and Emotional Behavior Risk Screener predicted academic proficiency at the end of the year (Kilgus, Bowman, Christ, & Taylor, 2017), the level of academic enablers as measured by the ESBA may predict which students would meet grade-level standards in language arts, math, and science. The SRSS predicts academic achievement and school discipline for European American and Latino Hispanic groups of students (Menzies & Lane, 2012). Additional studies could extend the predictive validity of the SRSS to African American, Asian American, and multiracial groups of students. Early detection of internalizing problems with SIBS provides an opportunity to change the course of the cascade social and emotional problems that cause personal distress and impair functioning at school (Masten & Cicchetti, 2010).

Future studies could examine the utility of the screeners to inform prevention and interventions in schools. At the classroom level, studies could explore how teachers use the information from universal screeners to adjust their instruction in social and
emotional learning in a manner similar to academic screening data (Deno & Mirkin, 1977; Fuchs, Deno, & Mirkin, 1984; Fuchs & Fuchs, 1986). At the school level, studies could examine how schools effectively use the information from aggregated data to focus interventions toward targeted groups of students who share common risk characteristics in a culturally responsive manner (McIntosh & Goodman, 2016). For example, when schools use universal screening data to measure the effects of their multi-tiered systems of support (Freeman, Miller, & Newcomer, 2015), future studies could examine the degree to which these prevention and intervention efforts reduce disproportionality in the referral to special education services and school disciplinary practices.

**Conclusion**

This study examined the construct validity of the *Elementary Social Behavior Assessment* (ESBA; Pennefather & Smolkowski, 2015), *Student Risk Screening Scale* (SRSS: Drummond, 1994), and the *Student Internalizing Behavior Screener* (SIBS; Cook et al., 2011). The results indicate that the total score from each screener holds the same meaning for African American, Asian American, Latino Hispanic, multiracial, and European American groups of students. The generalizability of the interpretation of the total scores applied to diverse racial and ethnic groups reduces bias in the making decisions regarding the identification of students at risk for surpassing the teacher’s threshold for teachability and developing externalizing or internalizing problems. Hopefully, these screeners contribute to the early identification of students in need of social, emotional, and behavior supports and the effective provision of interventions.
References


of early literacy skills in kindergarten with respect to English-language learners and ethnic subgroups. School Psychology Quarterly, 23(4), 553-570. doi: 10.1037/1045-3830.23.4.553


Bowman-Perrott, L., Burke, M. D., Zaini, S., Zhang, N., & Vannest, K. (2016). Promoting positive behavior using the Good Behavior Game: A meta-analysis of


professionals in dismantling the school to prison pipeline. *Journal of African American Males in Education, 1*(3), 196-211.


Eklund, K., & Dowdy, E. (2014). Screening for behavioral and emotional risk versus

Earhart, J. (2009). Early identification of behavioral and emotional problems in
School Psychologist, 14*(1), 89-95. doi: 10.1007/BF03340954

of academic enabling behaviors: An analysis of teachers’ and students’ ratings for

Elliot, S. N., & Gresham, F. M. (2008). *Social skills improvement system, performance
screening guide*. San Antonio, TX: PsychCorp Pearson Education.

Student Risk Screening Scale: A psychometrically sound, feasible tool for use in
urban elementary schools. *Journal of Emotional and Behavioral Disorders*, doi:
1063426611400082.


Horner, R. H., Todd, A. W., Lewis-Palmer, T., Irvin, L. K., Sugai, G., & Boland, J. B. (2004). The school-wide evaluation tool (SET) a research instrument for assessing...
school-wide positive behavior support. *Journal of Positive Behavior Interventions, 6*(1), 3-12. doi: [10.1177/10983007040060010201]


Keyes, C. L. (2009). The nature and importance of positive mental health in America’s adolescents. In Furlong, M. J., Gilman, R., & Huebner, E. S. (Eds.), *Handbook of positive psychology in schools* (pp. 9-23). New York: Routledge.


Lyon, A. R., Maras, M. A., Pate, C. M., Igusa, T., & Vander Stoep, A. (2016). Modeling the impact of school-based universal depression screening on additional service capacity needs: A system dynamics approach. *Administration and Policy in*
Mental Health and Mental Health Services Research, 43(2), 168-188. doi: 10.1007/s10488-015-0628-y


Marquez, B., Marquez, J., Vincent, C. G., Pennefather, J., Sprague, J. R., Smolkowski, K., & Yeaton, P. (2014). The iterative development and initial evaluation of we have skills! An innovative approach to teaching social skills to elementary students. Education and Treatment of Children, 37(1), 137-161. doi: 10.1353/etc.2014.0000


Simon, A. E., Pastor, P. N., Reuben, C. A., Huang, L. N., & Goldstrom, I. D. (2015). Use of mental health services by children ages six to 11 with emotional or behavioral


Chichester, UK: John Wiley & Sons Ltd. Retrieved from

http://faculty.educ.ubc.ca/zumbo/reprint_SEM_test_validation_Zumbo.pdf
Appendix A

Selection Process for Schools

Begin with the pool of elementary schools in the district.

Will the school reach 80-80 on the SET?  
No → Stop
Yes → Will the school complete the TFI?  
No → Stop
Yes → Will the school present an action plan?  
No → Stop
Yes → Will more than 70% of the students have fewer than 3 ODRs?  
No → Stop
Yes → Select the schools that meet all of the criteria and administer universal mental health screening.

SET = School-wide Evaluation Tool, TFI = Tiered Fidelity Inventory, ODR = Office discipline referrals.
Appendix B

Directions for Completing the Screeners

TOP Public Schools
2015 – 2016 Whole Child Snapshot Protocol

Whole Child Snapshot
When you go to the doctor they take your temperature, your blood pressure and weight. Anything that registers out of the norm (elevated blood pressure, unexplained weight gain/loss, etc.) can alert your doctor, who may investigate a little further. This information doesn’t tell the doctor what is wrong, just that something may be wrong.

The TWCI Snapshot operates under the same premises. It measures risk. It is a systematic and comprehensive way to examine schoolwide student behavioral needs and ensure students can be supported for academic success using a tiered system of supports. Teachers complete the snapshot three times a year (fall (6-8 wks after school starts), winter, spring). Collecting this information will provide the Tier 2/3 team with data to connect students to evidence-based interventions. The assessments used are highly predictive of student outcomes (office referrals, GPA, suspensions/expulsions), therefore critical to identify early and intervene and employ preventative strategies to improve student outcomes. There is little cost and time involved and the tools are research validated.

Steps:
1. Teachers meet together in a room with access to computers – each snapshot should take place at the same time of day to ensure the most accurate data.
2. Teachers log in to Performance Plus
3. Teachers complete ESBA, SRSS, and SIBS for every student in their class that has been in school for at least 30 days.
   - Teachers will score one student on all items before moving on to the next student (this prevents comparing students).
   - During snapshot:
      - follow your initial reaction – remember this is just a “quick snapshot”
      - Rate students independently without conferencing with other teachers or staff
      - limit talking to allow teachers to concentrate
      - no modification of the tools can occur – any modification will invalidate the supporting research.
      - there should be no operational definitions of the items, nor should staff develop any agreement on how to interpret the items.
4. Administration and Tier 2 team reviews data within a week and a half.

Reports
The reports available in Performance Plus are by class, grade and building level. The item level for each student should not be used to inform decisions – the overall score is the information that is used.
## Appendix C

### Items for Each Screener as Presented Online

<table>
<thead>
<tr>
<th>ESBA- Elementary School Behavior Assessment</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Listens to and respects the teacher</td>
<td></td>
</tr>
<tr>
<td>2. Follows the teacher’s directions</td>
<td></td>
</tr>
<tr>
<td>3. Works with effort</td>
<td></td>
</tr>
<tr>
<td>4. Does seatwork assignment as directed</td>
<td></td>
</tr>
<tr>
<td>5. Makes assistance needs known in appropriate manner</td>
<td></td>
</tr>
<tr>
<td>6. Follows rules</td>
<td></td>
</tr>
<tr>
<td>7. Avoids breaking rules even when encouraged by a peer</td>
<td></td>
</tr>
<tr>
<td>8. Behaves appropriately outside the classroom</td>
<td></td>
</tr>
<tr>
<td>9. Works out strong feelings appropriately</td>
<td></td>
</tr>
<tr>
<td>10. Can have normal conversations without becoming hostile</td>
<td></td>
</tr>
<tr>
<td>11. Gets along with peers</td>
<td></td>
</tr>
<tr>
<td>12. Resolves peer conflicts</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SRSS- Student Risk Screening Scale</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Steal</td>
<td></td>
</tr>
<tr>
<td>2. Lie, cheat, sneak</td>
<td></td>
</tr>
<tr>
<td>3. Behavior problems</td>
<td></td>
</tr>
<tr>
<td>4. Peer rejection</td>
<td></td>
</tr>
<tr>
<td>5. Low academic achievement</td>
<td></td>
</tr>
<tr>
<td>6. Negative attitude</td>
<td></td>
</tr>
<tr>
<td>7. Aggressive Behavior</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SIBS-Student Internalizing Behavior Screener</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Nervous or fearful</td>
<td></td>
</tr>
<tr>
<td>2. Bullied by peers</td>
<td></td>
</tr>
<tr>
<td>3. Spends time alone</td>
<td></td>
</tr>
<tr>
<td>4. Low academic achievement</td>
<td></td>
</tr>
<tr>
<td>5. Withdrawn</td>
<td></td>
</tr>
<tr>
<td>6. Sad or unhappy</td>
<td></td>
</tr>
<tr>
<td>7. Complains about being sick or hurt</td>
<td></td>
</tr>
</tbody>
</table>
Appendix D

Evaluating Model Fit and Comparison of Estimators of Measurement Invariance

Researchers encourage the presentation of multiple estimators for multigroup confirmatory factor analysis of measurement invariance using ordinal level data (DiStefano & Morgan, 2014; Lei, 2009; Li, 2016). The current study used a weighted least squares with means and variance adjustment (WLSMV) approach to address endogenous ordinal data (Rossel et al., 2017). For each research question, the screeners demonstrated adequate models with the CFI>.90 and the RMSEA<.08 using the various estimators. The ESBA uses the same respecified model for all of the estimators. The multigroup confirmatory factor analysis based on each estimator produced trivial changes in the CFI as constraints increased, illustrating the robustness of measurement invariance based on various estimators.

The estimation methods address differently categorical data and non-normal distributions. Each method of estimation produces different levels of bias from the high degree kurtosis. The robust methods use the scaling correction factor to adjust the means and standard errors for the their test statistic, namely the weighted least squares with adjusted means (WLSM), the weighted least squares with adjusted means and variances (WLSMV), the unweighted least squares with adjusted means (ULSM), and the maximum likelihood (MLR).

The WLS methods do not assume any distributional form by using an arbitrary distribution function to include the degree of skew and kurtosis in the estimate. Whereas the ULS method uses the entire covariances of covariance matrix, the WLS methods uses a computationally simpler approach based on the weighted diagonal of the covariances of
covariance matrix. The weighted least squares with means and variances (WLSMV) typically produces the least Type I errors with ordinal data (Sass, Schmidt, & March, 2014). The chi square tests of absolute fit are overly sensitive to large sample size and non-normality of the distribution.

Comparing the various methods of estimation at each level of measurement invariance involves evaluating the overlap of the confidence intervals from the RMSEA and also the change in CFI as discussed previously in Chapter 3. The nested models that do not differ when the RMSEA values fall within one another’s confidence intervals indicate measurement invariance. The change in CFI between nested models that falls below positive .01 provides evidence of measurement invariance.

The maximum likelihood (MLR) method permits using a scaled difference in chi square test for nested models (Satorra & Bentler, 2001) to determine measurement invariance. The scaled difference test accounts for the non-normal chi square distributions used the nested models (Yuan & Bentler, 2000). The scaled difference in chi square test for nested models additionally tests for change in CFI for measurement invariance.

For research question one, each of the models of the screeners demonstrates configural invariance according to the various estimators with the CFI>.90 and the RMSEA<.08 (Table 13). The 90% confidence intervals for the RMSEA on both the ESBA and the SRSS overlap each of their respective estimators. On the SIBS, the confidence intervals of the RMSEA from the WLSMV and the MLR fall outside of range of the other. Apart from these estimation methods on the SIBS, the models from the
remaining estimation models demonstrate configural invariance on the ESBA and the SRSS.

Table 13
*Configural Invariance Estimated with WLSMV, WLSM, ULSM, and MLR*

<table>
<thead>
<tr>
<th>Estimator</th>
<th>$\chi^2$ (df)</th>
<th>RMSEA</th>
<th>P-value</th>
<th>CFI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ESBA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WLSMV</td>
<td>651.080 (210)</td>
<td>0.052 (0.047, 0.056)</td>
<td>0.263</td>
<td>0.928</td>
</tr>
<tr>
<td>WLSM</td>
<td>865.719 (210)</td>
<td>0.063 (0.047, 0.080)</td>
<td>0.088</td>
<td>0.982</td>
</tr>
<tr>
<td>ULSM</td>
<td>851.530 (210)</td>
<td>0.062 (0.041, 0.085)</td>
<td>0.158</td>
<td>0.983</td>
</tr>
<tr>
<td>MLR</td>
<td>637.191 (210)</td>
<td>0.051 (0.048, 0.054)</td>
<td>0.333</td>
<td>0.979</td>
</tr>
</tbody>
</table>

| **SRSS**  |               |       |        |     |
| WLSMV     | 256.245 (70)  | 0.058 (0.051, 0.066)$^a$ | 0.034 | 0.947 |
| WLSM      | 286.284 (70)  | 0.063 (0.047, 0.080) | 0.085 | 0.975 |
| ULSM      | 290.201 (70)  | 0.064 (0.048, 0.080) | 0.074 | 0.979 |
| MLR       | 351.550 (70)  | 0.072 (0.067, 0.077)$^a$ | 0.000 | 0.953$^g$ |

| **SIBS**  |               |       |        |     |
| WLSMV     | 190.867 (70)  | 0.047 (0.039, 0.055) | 0.729 | 0.945 |
| WLSM      | 207.239 (70)  | 0.050 (0.036, 0.064) | 0.477 | 0.971 |
| ULSM      | 215.960 (70)  | 0.052 (0.036, 0.067) | 0.409 | 0.973 |
| MLR       | 285.820 (70)  | 0.063 (0.057, 0.068)$^a$ | 0.000 | 0.941 |

* $p = .001$ ** $p < .001$ $^g$ = robust $^a$ = upper bound of 90% confidence interval < .08
For research question two, each of the screeners demonstrates adequate models with constrained factor loadings according to the various estimators with the CFI>.90 and the RMSEA<.08 (Table 14). On the ESBA, the SRSS, and the SIBS the 90% confidence intervals do not overlap between the WLSMV and the MLR estimators. Yet, the change in CFI indicates that the adequacy of the models improves as the constraints on the loadings are fixed to the European American group, except with the MLR estimator for the ESBA and the SRSS. Using the Satorra and Bentler (2001) scaled difference in chi square test for nested models based on the MLR estimator, the metric CFI differs from the configural CFI by .002 on the ESBA, .007 on the SRSS, and .003 on the SIBS. None of the changes in CFI from the configural to metric models degrade the models significantly, indicating metric invariance across all estimators.
Table 14

Metric Invariance Estimated with WLSMV, WLSM, ULSM, and MLR

<table>
<thead>
<tr>
<th>Estimator</th>
<th>$\chi^2$ (df)</th>
<th>RMSEA</th>
<th>P-value RMSEA &lt;= 0.05</th>
<th>CFI</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESBA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WLSMV</td>
<td>651.080 (210)</td>
<td>0.052</td>
<td>0.263</td>
<td>0.928</td>
</tr>
<tr>
<td></td>
<td>865.719 (210)</td>
<td>0.063</td>
<td>0.088</td>
<td>0.982</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WLSM</td>
<td>851.530 (210)</td>
<td>0.062</td>
<td>0.158</td>
<td>0.983</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ULSM</td>
<td>637.191 (210)</td>
<td>0.051</td>
<td>0.333</td>
<td>0.979</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLR</td>
<td>256.245 (70)</td>
<td>0.058</td>
<td>0.034</td>
<td>0.947</td>
</tr>
<tr>
<td></td>
<td>286.284 (70)</td>
<td>0.063</td>
<td>0.085</td>
<td>0.975</td>
</tr>
<tr>
<td></td>
<td>290.201 (70)</td>
<td>0.064</td>
<td>0.074</td>
<td>0.979</td>
</tr>
<tr>
<td></td>
<td>351.550 (70)</td>
<td>0.072</td>
<td>0.000</td>
<td>0.951</td>
</tr>
<tr>
<td>SRSS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WLSMV</td>
<td>190.867 (70)</td>
<td>0.047</td>
<td>0.729</td>
<td>0.945</td>
</tr>
<tr>
<td></td>
<td>207.239 (70)</td>
<td>0.050</td>
<td>0.477</td>
<td>0.971</td>
</tr>
<tr>
<td></td>
<td>215.960 (70)</td>
<td>0.052</td>
<td>0.409</td>
<td>0.973</td>
</tr>
<tr>
<td></td>
<td>285.820 (70)</td>
<td>0.063</td>
<td>0.000</td>
<td>0.941</td>
</tr>
<tr>
<td>SIBS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WLSMV</td>
<td>190.867 (70)</td>
<td>0.047</td>
<td>0.729</td>
<td>0.945</td>
</tr>
<tr>
<td></td>
<td>207.239 (70)</td>
<td>0.050</td>
<td>0.477</td>
<td>0.971</td>
</tr>
<tr>
<td></td>
<td>215.960 (70)</td>
<td>0.052</td>
<td>0.409</td>
<td>0.973</td>
</tr>
<tr>
<td></td>
<td>285.820 (70)</td>
<td>0.063</td>
<td>0.000</td>
<td>0.941</td>
</tr>
</tbody>
</table>

* $p = .001$ ** $p < .001$  § = robust  a = upper bound of 90% confidence interval < .08

For research question three, each of the screeners demonstrates adequate models with constrained thresholds and factor loadings according to the various estimators with
the CFI>.90 and the RMSEA<.08 (Table 15). On the ESBA, the SRSS, and the SIBS, the 90% confidence intervals do not overlap between the WLSMV and the MLR estimators. Moreover, the lack of change in the CFI indicates that the metric and scalar models do not differ for the ESBA across all estimators and similarly the WLSMV and ULSM on the SRSS, and the WLSM for the SIBS. The model negligibly improved for the WLSMV on the SIBS. Otherwise, the change in CFI barely diminished on the WLSM for the SRSS and on the ULSM for the SIBS. The scaled difference in chi square test for nested models based on the MLR estimator indicates that the scalar CFI differs from the metric CFI by .001 on the ESBA, .007 on the SRSS, and .019 on the SIBS. Thus, only the SIBS fails to demonstrate metric invariance based on all of the estimation methods due to a high scaled difference score comparing nested models based on MLR.
Table 15
Scalar Invariance Estimated with WLSMV, WLSM, ULSM, and MLR

<table>
<thead>
<tr>
<th>Estimator</th>
<th>$\chi^2$ (df)</th>
<th>RMSEA</th>
<th>P-value</th>
<th>CFI</th>
<th>$\Delta$ CFI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESBA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WLSMV</td>
<td>449.129 (250)**</td>
<td>0.032</td>
<td>1.000</td>
<td>0.967</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>643.478 (250)**</td>
<td>0.045</td>
<td>0.826</td>
<td>0.989</td>
<td>.000</td>
</tr>
<tr>
<td>WLSM</td>
<td>658.404 (250)**</td>
<td>0.046</td>
<td>0.714</td>
<td>0.989</td>
<td>.000</td>
</tr>
<tr>
<td>ULSM</td>
<td>698.272 (250)**</td>
<td>0.048</td>
<td>0.890</td>
<td>0.978</td>
<td>.000</td>
</tr>
<tr>
<td>MLR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRSS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WLSMV</td>
<td>190.861 (90)**</td>
<td>0.038</td>
<td>0.997</td>
<td>0.971</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>238.817 (90)**</td>
<td>0.046</td>
<td>0.731</td>
<td>0.983</td>
<td>.001</td>
</tr>
<tr>
<td>WLSM</td>
<td>232.169 (90)**</td>
<td>0.045</td>
<td>0.792</td>
<td>0.986</td>
<td>.000</td>
</tr>
<tr>
<td>ULSM</td>
<td>369.520 (90)**</td>
<td>0.063</td>
<td>0.000</td>
<td>0.951</td>
<td>.002</td>
</tr>
<tr>
<td>MLR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIBS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WLSMV</td>
<td>136.743 (90)*</td>
<td>0.026</td>
<td>1.000</td>
<td>0.979</td>
<td>-.001</td>
</tr>
<tr>
<td></td>
<td>150.322 (90)*</td>
<td>0.029</td>
<td>1.000</td>
<td>0.987</td>
<td>.000</td>
</tr>
<tr>
<td>WLSM</td>
<td>160.892 (90)**</td>
<td>0.032</td>
<td>0.997</td>
<td>0.987</td>
<td>.001</td>
</tr>
<tr>
<td>MLR</td>
<td>278.448 (90)**</td>
<td>0.052</td>
<td>0.285</td>
<td>0.949</td>
<td>.003</td>
</tr>
</tbody>
</table>

* $p = .001$ ** $p < .001$ § = robust  a = upper bound of 90% confidence interval < .08
Appendix E

Computer Application for Data Analysis

The R statistical applications for statistical computing and graphics will analyze the data (R Core Team, 2016). The R application is freely available, open source software that uses packages to perform the data manipulation, calculation and graphical display for this study. This study uses the psych, lavaan, MVM, qgraph, semPlot, and semTools packages. The R packages perform the functions found in the popular structural equation modeling software mPlus (Muthén & Muthén, 1998-2014). The psych package performs psychometric functions (Revelle, 2015). The lavaan package estimates latent variable models such as structural equation models and multi-group confirmatory factor analysis (Rosseel, 2012; Rosseel et al., 2017). The semPlot package constructs path diagrams for visual analysis of parameter estimates (Epskamp, 2015). The qgraph package visualizes correlation matrices for structural equation models (Epskamp, Cramer, Waldorp, Schmittmann, & Borsboom, 2012).
Appendix F

Polychoric Correlation Matrix

The polychoric correlation matrix shows the correlation between items based on ordinal data. As a result of using thresholds rather than means, the polychoric correlation matrix does not include standard deviations from the mean. The items on the ESBA correlate from 0.511 to 0.833 (Table 16). All of the items on the SRSS correlate between 0.320 to 0.718, with the exception of two pairs below 0.30 (Table 17). Many of the items on the SIBS correlate between 0.312 to 0.599, except for six pairs below 0.30 (Table 18).

Table 16
Polychoric Correlation Matrix for ESBA

<table>
<thead>
<tr>
<th>Item</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>T5</th>
<th>T6</th>
<th>T7</th>
<th>T8</th>
<th>T9</th>
<th>T10</th>
<th>T11</th>
<th>T12</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2</td>
<td>0.822</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T3</td>
<td>0.644</td>
<td>0.707</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T4</td>
<td>0.650</td>
<td>0.724</td>
<td>0.833</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T5</td>
<td>0.631</td>
<td>0.658</td>
<td>0.685</td>
<td>0.707</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T6</td>
<td>0.786</td>
<td>0.791</td>
<td>0.650</td>
<td>0.673</td>
<td>0.643</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T7</td>
<td>0.697</td>
<td>0.693</td>
<td>0.594</td>
<td>0.605</td>
<td>0.600</td>
<td>0.766</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T8</td>
<td>0.711</td>
<td>0.704</td>
<td>0.588</td>
<td>0.606</td>
<td>0.601</td>
<td>0.765</td>
<td>0.742</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T9</td>
<td>0.657</td>
<td>0.642</td>
<td>0.557</td>
<td>0.583</td>
<td>0.645</td>
<td>0.662</td>
<td>0.639</td>
<td>0.710</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T10</td>
<td>0.653</td>
<td>0.621</td>
<td>0.511</td>
<td>0.535</td>
<td>0.575</td>
<td>0.655</td>
<td>0.603</td>
<td>0.685</td>
<td>0.724</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T11</td>
<td>0.639</td>
<td>0.632</td>
<td>0.539</td>
<td>0.586</td>
<td>0.693</td>
<td>0.618</td>
<td>0.727</td>
<td>0.718</td>
<td>0.719</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T12</td>
<td>0.625</td>
<td>0.629</td>
<td>0.537</td>
<td>0.568</td>
<td>0.609</td>
<td>0.648</td>
<td>0.659</td>
<td>0.693</td>
<td>0.747</td>
<td>0.680</td>
<td>0.760</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Note: E1-E5 and E5-E7 show polychoric correlations below .30.

Table 17
Polychoric Correlation Matrix for SRSS

<table>
<thead>
<tr>
<th>Item</th>
<th>E1</th>
<th>E2</th>
<th>E3</th>
<th>E4</th>
<th>E5</th>
<th>E6</th>
<th>E7</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E2</td>
<td>0.581</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E3</td>
<td>0.450</td>
<td>0.664</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E4</td>
<td>0.399</td>
<td>0.499</td>
<td>0.565</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E5</td>
<td>0.216</td>
<td>0.313</td>
<td>0.372</td>
<td>0.320</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E6</td>
<td>0.392</td>
<td>0.571</td>
<td>0.718</td>
<td>0.569</td>
<td>0.349</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>E7</td>
<td>0.434</td>
<td>0.551</td>
<td>0.693</td>
<td>0.534</td>
<td>0.275</td>
<td>0.645</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Table 18

*Polychoric Correlation Matrix for SIBS*

<table>
<thead>
<tr>
<th>Item</th>
<th>I1</th>
<th>I2</th>
<th>I3</th>
<th>I4</th>
<th>I5</th>
<th>I6</th>
<th>I7</th>
</tr>
</thead>
<tbody>
<tr>
<td>I1</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I2</td>
<td>0.379</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I3</td>
<td>0.419</td>
<td>0.396</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I4</td>
<td>0.280</td>
<td>0.239</td>
<td>0.271</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I5</td>
<td>0.460</td>
<td>0.351</td>
<td>0.599</td>
<td>0.327</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I6</td>
<td>0.498</td>
<td>0.409</td>
<td>0.474</td>
<td>0.308</td>
<td>0.580</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>I7</td>
<td>0.333</td>
<td>0.289</td>
<td>0.260</td>
<td>0.203</td>
<td>0.312</td>
<td>0.452</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Note: I1-I4, I2-I4, I2-I7, I3-I4, I3-I7, and I4-I7 show polychoric correlations below .30.
Appendix G

Polychoric Residuals for the ESBA

Table 19
Polychoric Residuals for the ESBA

<table>
<thead>
<tr>
<th></th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>T5</th>
<th>T6</th>
<th>T7</th>
<th>T8</th>
<th>T9</th>
<th>T10</th>
<th>T11</th>
<th>T12</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2</td>
<td>0.097</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T3</td>
<td>0.003</td>
<td>0.056</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T4</td>
<td>-0.012</td>
<td>0.051</td>
<td>0.238</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T5</td>
<td>-0.024</td>
<td>-0.008</td>
<td>0.097</td>
<td>0.099</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T6</td>
<td>0.052</td>
<td>0.044</td>
<td>-0.010</td>
<td>-0.009</td>
<td>-0.031</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T7</td>
<td>0.014</td>
<td>-0.002</td>
<td>-0.020</td>
<td>-0.030</td>
<td>-0.028</td>
<td>0.062</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T8</td>
<td>-0.003</td>
<td>-0.022</td>
<td>-0.054</td>
<td>-0.058</td>
<td>-0.055</td>
<td>0.030</td>
<td>0.057</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T9</td>
<td>-0.028</td>
<td>-0.054</td>
<td>-0.058</td>
<td>-0.053</td>
<td>0.015</td>
<td>-0.043</td>
<td>-0.017</td>
<td>0.024</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T10</td>
<td>0.000</td>
<td>-0.044</td>
<td>-0.075</td>
<td>-0.071</td>
<td>-0.025</td>
<td>-0.018</td>
<td>-0.024</td>
<td>0.031</td>
<td>0.096</td>
<td>0.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T11</td>
<td>-0.035</td>
<td>-0.053</td>
<td>-0.066</td>
<td>-0.060</td>
<td>-0.036</td>
<td>-0.024</td>
<td>-0.028</td>
<td>0.052</td>
<td>0.071</td>
<td>0.102</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>T12</td>
<td>-0.045</td>
<td>-0.053</td>
<td>-0.065</td>
<td>-0.055</td>
<td>-0.007</td>
<td>-0.042</td>
<td>0.016</td>
<td>0.021</td>
<td>0.102</td>
<td>0.066</td>
<td>0.126</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Note: T3-T4, T9-T12, T10-T11, T11-T12 present elevated polychoric residuals.
Figure 8 illustrates the residuals from the polychoric correlation matrix from Table 19. The wider lines between two items convey the magnitude of mismatch between the implied model and the observed model. The green lines connecting suggest that the items share.

Figure 8. Network of polychoric residuals.

Note: The red lines indicate that the implied model over predicted the observed model. The green lines indicate that the implied model underpredicted the observed model. The width of the line corresponds to the degree of mismatch between the implied and observed models.
Appendix H

Parameter Estimates for Scalar Invariance on the ESBA by Racial Ethnic Group

Table 20
Parameter Estimates for Scalar Invariance on the ESBA for African American Students.

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Unstandardized Loadings</th>
<th>Standard Error</th>
<th>Standardized Loadings</th>
<th>Unstandardized Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>1.000</td>
<td></td>
<td>0.833</td>
<td>3.688</td>
</tr>
<tr>
<td>T2</td>
<td>1.024</td>
<td>0.015</td>
<td>0.803</td>
<td>3.629</td>
</tr>
<tr>
<td>T3</td>
<td>0.937</td>
<td>0.024</td>
<td>0.669</td>
<td>3.598</td>
</tr>
<tr>
<td>T4</td>
<td>0.955</td>
<td>0.023</td>
<td>0.694</td>
<td>3.601</td>
</tr>
<tr>
<td>T5</td>
<td>1.050</td>
<td>0.023</td>
<td>0.723</td>
<td>3.589</td>
</tr>
<tr>
<td>T6</td>
<td>1.058</td>
<td>0.018</td>
<td>0.884</td>
<td>3.660</td>
</tr>
<tr>
<td>T7</td>
<td>1.144</td>
<td>0.024</td>
<td>0.839</td>
<td>3.584</td>
</tr>
<tr>
<td>T8</td>
<td>1.055</td>
<td>0.021</td>
<td>0.840</td>
<td>3.669</td>
</tr>
<tr>
<td>T9</td>
<td>1.118</td>
<td>0.024</td>
<td>0.741</td>
<td>3.547</td>
</tr>
<tr>
<td>T10</td>
<td>0.823</td>
<td>0.023</td>
<td>0.736</td>
<td>3.760</td>
</tr>
<tr>
<td>T11</td>
<td>0.851</td>
<td>0.023</td>
<td>0.721</td>
<td>3.691</td>
</tr>
<tr>
<td>T12</td>
<td>1.096</td>
<td>0.027</td>
<td>0.720</td>
<td>3.541</td>
</tr>
</tbody>
</table>

Figure 9. Scalar model of the ESBA for African American students.
Table 21
Parameter Estimates for Scalar Invariance on the ESBA for Asian American Students.

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Unstandardized Loadings</th>
<th>Standard Error</th>
<th>Standardized Loadings</th>
<th>Unstandardized Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>1.000</td>
<td>0.864</td>
<td></td>
<td>3.571</td>
</tr>
<tr>
<td>T2</td>
<td>1.024</td>
<td>0.015</td>
<td>0.842</td>
<td>3.502</td>
</tr>
<tr>
<td>T3</td>
<td>0.937</td>
<td>0.024</td>
<td>0.782</td>
<td>3.523</td>
</tr>
<tr>
<td>T4</td>
<td>0.955</td>
<td>0.023</td>
<td>0.810</td>
<td>3.541</td>
</tr>
<tr>
<td>T5</td>
<td>1.050</td>
<td>0.023</td>
<td>0.809</td>
<td>3.491</td>
</tr>
<tr>
<td>T6</td>
<td>1.058</td>
<td>0.018</td>
<td>0.883</td>
<td>3.523</td>
</tr>
<tr>
<td>T7</td>
<td>1.144</td>
<td>0.024</td>
<td>0.860</td>
<td>3.406</td>
</tr>
<tr>
<td>T8</td>
<td>1.055</td>
<td>0.021</td>
<td>0.856</td>
<td>3.521</td>
</tr>
<tr>
<td>T9</td>
<td>1.118</td>
<td>0.024</td>
<td>0.806</td>
<td>3.416</td>
</tr>
<tr>
<td>T10</td>
<td>0.823</td>
<td>0.023</td>
<td>0.728</td>
<td>3.685</td>
</tr>
<tr>
<td>T11</td>
<td>0.851</td>
<td>0.023</td>
<td>0.765</td>
<td>3.584</td>
</tr>
<tr>
<td>T12</td>
<td>1.096</td>
<td>0.027</td>
<td>0.792</td>
<td>3.402</td>
</tr>
</tbody>
</table>

Figure 10. Scalar model of the ESBA for Asian American students.
Table 22
Parameter Estimates for Scalar Invariance on the ESBA for Latino Hispanic Students.

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Unstandardized Loadings</th>
<th>Standard Error</th>
<th>Standardized Loadings</th>
<th>Unstandardized Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>1.000</td>
<td>0.855</td>
<td></td>
<td>3.702</td>
</tr>
<tr>
<td>T2</td>
<td>1.024</td>
<td>0.843</td>
<td></td>
<td>3.664</td>
</tr>
<tr>
<td>T3</td>
<td>0.937</td>
<td>0.676</td>
<td></td>
<td>3.570</td>
</tr>
<tr>
<td>T4</td>
<td>0.955</td>
<td>0.700</td>
<td></td>
<td>3.596</td>
</tr>
<tr>
<td>T5</td>
<td>1.050</td>
<td>0.719</td>
<td></td>
<td>3.551</td>
</tr>
<tr>
<td>T6</td>
<td>1.058</td>
<td>0.895</td>
<td></td>
<td>3.687</td>
</tr>
<tr>
<td>T7</td>
<td>1.144</td>
<td>0.808</td>
<td></td>
<td>3.573</td>
</tr>
<tr>
<td>T8</td>
<td>1.055</td>
<td>0.901</td>
<td></td>
<td>3.723</td>
</tr>
<tr>
<td>T9</td>
<td>1.118</td>
<td>0.844</td>
<td></td>
<td>3.641</td>
</tr>
<tr>
<td>T10</td>
<td>0.823</td>
<td>0.799</td>
<td></td>
<td>3.819</td>
</tr>
<tr>
<td>T11</td>
<td>0.851</td>
<td>0.800</td>
<td></td>
<td>3.763</td>
</tr>
<tr>
<td>T12</td>
<td>1.096</td>
<td>0.773</td>
<td></td>
<td>3.594</td>
</tr>
</tbody>
</table>

Figure 11. Scalar model of the ESBA for Latino Hispanic students.
Table 23
*Parameter Estimates for Scalar Invariance on the ESBA for Multiracial Students*

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Unstandardized Loadings</th>
<th>Standard Error</th>
<th>Standardized Loadings</th>
<th>Unstandardized Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>1.000</td>
<td>0.869</td>
<td>3.784</td>
<td></td>
</tr>
<tr>
<td>T2</td>
<td>1.024</td>
<td>0.015</td>
<td>0.834</td>
<td>3.752</td>
</tr>
<tr>
<td>T3</td>
<td>0.937</td>
<td>0.024</td>
<td>0.661</td>
<td>3.718</td>
</tr>
<tr>
<td>T4</td>
<td>0.955</td>
<td>0.023</td>
<td>0.683</td>
<td>3.729</td>
</tr>
<tr>
<td>T5</td>
<td>1.050</td>
<td>0.023</td>
<td>0.733</td>
<td>3.700</td>
</tr>
<tr>
<td>T6</td>
<td>1.058</td>
<td>0.018</td>
<td>0.886</td>
<td>3.778</td>
</tr>
<tr>
<td>T7</td>
<td>1.144</td>
<td>0.024</td>
<td>0.751</td>
<td>3.664</td>
</tr>
<tr>
<td>T8</td>
<td>1.055</td>
<td>0.021</td>
<td>0.888</td>
<td>3.789</td>
</tr>
<tr>
<td>T9</td>
<td>1.118</td>
<td>0.024</td>
<td>0.781</td>
<td>3.728</td>
</tr>
<tr>
<td>T10</td>
<td>0.823</td>
<td>0.023</td>
<td>0.838</td>
<td>3.882</td>
</tr>
<tr>
<td>T11</td>
<td>0.851</td>
<td>0.023</td>
<td>0.729</td>
<td>3.810</td>
</tr>
<tr>
<td>T12</td>
<td>1.096</td>
<td>0.027</td>
<td>0.750</td>
<td>3.726</td>
</tr>
</tbody>
</table>

*Figure 12.* Scalar model of the ESBA for multiracial students.
Table 24
Parameter Estimates for Scalar Invariance on the ESBA for European American Students

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Unstandardized Loadings</th>
<th>Standard Error</th>
<th>Standardized Loadings</th>
<th>Unstandardized Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>1.000</td>
<td></td>
<td>0.832</td>
<td>3.472</td>
</tr>
<tr>
<td>T2</td>
<td>1.024</td>
<td>0.015</td>
<td>0.841</td>
<td>3.399</td>
</tr>
<tr>
<td>T3</td>
<td>0.937</td>
<td>0.024</td>
<td>0.738</td>
<td>3.411</td>
</tr>
<tr>
<td>T4</td>
<td>0.955</td>
<td>0.023</td>
<td>0.757</td>
<td>3.424</td>
</tr>
<tr>
<td>T5</td>
<td>1.050</td>
<td>0.023</td>
<td>0.786</td>
<td>3.386</td>
</tr>
<tr>
<td>T6</td>
<td>1.058</td>
<td>0.018</td>
<td>0.884</td>
<td>3.450</td>
</tr>
<tr>
<td>T7</td>
<td>1.144</td>
<td>0.024</td>
<td>0.812</td>
<td>3.301</td>
</tr>
<tr>
<td>T8</td>
<td>1.055</td>
<td>0.021</td>
<td>0.859</td>
<td>3.460</td>
</tr>
<tr>
<td>T9</td>
<td>1.118</td>
<td>0.024</td>
<td>0.773</td>
<td>3.317</td>
</tr>
<tr>
<td>T10</td>
<td>0.823</td>
<td>0.023</td>
<td>0.696</td>
<td>3.571</td>
</tr>
<tr>
<td>T11</td>
<td>0.851</td>
<td>0.023</td>
<td>0.779</td>
<td>3.528</td>
</tr>
<tr>
<td>T12</td>
<td>1.096</td>
<td>0.027</td>
<td>0.778</td>
<td>3.294</td>
</tr>
</tbody>
</table>

Figure 13. Scalar model of the ESBA for European American students.
## Appendix I

Parameter Estimates for Scalar Invariance on the SRSS by Racial Ethnic Group

Table 25

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Unstandardized Loadings</th>
<th>Standard Error</th>
<th>Standardized Loadings</th>
<th>Unstandardized Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>E3</td>
<td>1.000</td>
<td>0.881</td>
<td>1.700</td>
<td></td>
</tr>
<tr>
<td>E1</td>
<td>0.301</td>
<td>0.017</td>
<td>0.569</td>
<td>1.145</td>
</tr>
<tr>
<td>E2</td>
<td>0.704</td>
<td>0.018</td>
<td>0.748</td>
<td>1.471</td>
</tr>
<tr>
<td>E4</td>
<td>0.584</td>
<td>0.018</td>
<td>0.651</td>
<td>1.421</td>
</tr>
<tr>
<td>E5</td>
<td>0.534</td>
<td>0.022</td>
<td>0.450</td>
<td>1.776</td>
</tr>
<tr>
<td>E6</td>
<td>0.827</td>
<td>0.016</td>
<td>0.816</td>
<td>1.547</td>
</tr>
<tr>
<td>E7</td>
<td>0.657</td>
<td>0.017</td>
<td>0.750</td>
<td>1.361</td>
</tr>
</tbody>
</table>

Figure 14. Scalar model of the SRSS for African American students.
Table 26
Parameter Estimates for Scalar Invariance on the SRSS for Asian American Students.

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Unstandardized Loadings</th>
<th>Standard Error</th>
<th>Standardized Loadings</th>
<th>Unstandardized Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>E3</td>
<td>1.000</td>
<td>0.017</td>
<td>0.898</td>
<td>1.902</td>
</tr>
<tr>
<td>E1</td>
<td>0.301</td>
<td>0.017</td>
<td>0.519</td>
<td>1.212</td>
</tr>
<tr>
<td>E2</td>
<td>0.704</td>
<td>0.018</td>
<td>0.775</td>
<td>1.656</td>
</tr>
<tr>
<td>E4</td>
<td>0.584</td>
<td>0.018</td>
<td>0.704</td>
<td>1.484</td>
</tr>
<tr>
<td>E5</td>
<td>0.534</td>
<td>0.022</td>
<td>0.466</td>
<td>1.926</td>
</tr>
<tr>
<td>E6</td>
<td>0.827</td>
<td>0.016</td>
<td>0.817</td>
<td>1.735</td>
</tr>
<tr>
<td>E7</td>
<td>0.657</td>
<td>0.017</td>
<td>0.754</td>
<td>1.479</td>
</tr>
</tbody>
</table>

Figure 15. Scalar model of the SRSS for Asian American students.
Table 27
*Parameter Estimates for Scalar Invariance on the SRSS for Latino Hispanic Students*

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Unstandardized Loadings</th>
<th>Standard Error</th>
<th>Standardized Loadings</th>
<th>Unstandardized Loadings</th>
<th>Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>E3</td>
<td>1.000</td>
<td></td>
<td></td>
<td>0.836</td>
<td>1.524</td>
</tr>
<tr>
<td>E1</td>
<td>0.301</td>
<td>0.017</td>
<td>0.544</td>
<td>1.101</td>
<td></td>
</tr>
<tr>
<td>E2</td>
<td>0.704</td>
<td>0.018</td>
<td>0.711</td>
<td>1.381</td>
<td></td>
</tr>
<tr>
<td>E4</td>
<td>0.584</td>
<td>0.018</td>
<td>0.644</td>
<td>1.271</td>
<td></td>
</tr>
<tr>
<td>E5</td>
<td>0.534</td>
<td>0.022</td>
<td>0.349</td>
<td>2.010</td>
<td></td>
</tr>
<tr>
<td>E6</td>
<td>0.827</td>
<td>0.016</td>
<td>0.764</td>
<td>1.402</td>
<td></td>
</tr>
<tr>
<td>E7</td>
<td>0.657</td>
<td>0.017</td>
<td>0.806</td>
<td>1.219</td>
<td></td>
</tr>
</tbody>
</table>

*Figure 16.* Scalar model of the SRSS for Latino Hispanic students.
Table 28

**Parameter Estimates for Scalar Invariance on the SRSS for Multiracial Students**

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Unstandardized Loadings</th>
<th>Standard Error</th>
<th>Standardized Loadings</th>
<th>Unstandardized Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>E3</td>
<td>1.000</td>
<td>0.751</td>
<td>1.391</td>
<td></td>
</tr>
<tr>
<td>E1</td>
<td>0.301</td>
<td>0.017</td>
<td>0.691</td>
<td>1.053</td>
</tr>
<tr>
<td>E2</td>
<td>0.704</td>
<td>0.018</td>
<td>0.667</td>
<td>1.272</td>
</tr>
<tr>
<td>E4</td>
<td>0.584</td>
<td>0.018</td>
<td>0.652</td>
<td>1.181</td>
</tr>
<tr>
<td>E5</td>
<td>0.534</td>
<td>0.022</td>
<td>0.293</td>
<td>1.755</td>
</tr>
<tr>
<td>E6</td>
<td>0.827</td>
<td>0.016</td>
<td>0.784</td>
<td>1.234</td>
</tr>
<tr>
<td>E7</td>
<td>0.657</td>
<td>0.017</td>
<td>0.777</td>
<td>1.151</td>
</tr>
</tbody>
</table>

*Figure 17. Scalar model of the SRSS for multiracial students.*
Table 29
Parameter Estimates for Scalar Invariance on the SRSS for European American Students

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Unstandardized Loadings</th>
<th>Standard Error</th>
<th>Standardized Loadings</th>
<th>Unstandardized Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>E3</td>
<td>1.000</td>
<td>0.017</td>
<td>0.885</td>
<td>2.031</td>
</tr>
<tr>
<td>E1</td>
<td>0.301</td>
<td>0.017</td>
<td>0.500</td>
<td>1.239</td>
</tr>
<tr>
<td>E2</td>
<td>0.704</td>
<td>0.018</td>
<td>0.748</td>
<td>1.700</td>
</tr>
<tr>
<td>E4</td>
<td>0.584</td>
<td>0.018</td>
<td>0.715</td>
<td>1.494</td>
</tr>
<tr>
<td>E5</td>
<td>0.534</td>
<td>0.022</td>
<td>0.458</td>
<td>2.177</td>
</tr>
<tr>
<td>E6</td>
<td>0.827</td>
<td>0.016</td>
<td>0.785</td>
<td>1.922</td>
</tr>
<tr>
<td>E7</td>
<td>0.657</td>
<td>0.017</td>
<td>0.751</td>
<td>1.518</td>
</tr>
</tbody>
</table>

Figure 18. Scalar model of the SRSS for European American students.
Appendix J

Parameter Estimates for Scalar Invariance on the SIBS by Racial Ethnic Group

Table 30
Parameter Estimates for Scalar Invariance on the SIBS for African American Students

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Unstandardized Loadings</th>
<th>Standard Error</th>
<th>Standardized Loadings</th>
<th>Unstandardized Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>I1</td>
<td>1.000</td>
<td></td>
<td>0.654</td>
<td>1.396</td>
</tr>
<tr>
<td>I2</td>
<td>0.626</td>
<td>0.035</td>
<td>0.523</td>
<td>1.249</td>
</tr>
<tr>
<td>I3</td>
<td>1.021</td>
<td>0.045</td>
<td>0.661</td>
<td>1.408</td>
</tr>
<tr>
<td>I4</td>
<td>1.002</td>
<td>0.053</td>
<td>0.475</td>
<td>1.754</td>
</tr>
<tr>
<td>I5</td>
<td>1.100</td>
<td>0.047</td>
<td>0.794</td>
<td>1.333</td>
</tr>
<tr>
<td>I6</td>
<td>1.297</td>
<td>0.052</td>
<td>0.816</td>
<td>1.463</td>
</tr>
<tr>
<td>I7</td>
<td>0.720</td>
<td>0.041</td>
<td>0.485</td>
<td>1.358</td>
</tr>
</tbody>
</table>

Figure 19. Scalar model of the SIBS for African American students.
**Table 31**  
*Parameter Estimates for Scalar Invariance on the SIBS for Asian American Students*

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Unstandardized Loadings</th>
<th>Standard Error</th>
<th>Standardized Loadings</th>
<th>Unstandardized Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>I1</td>
<td>1.000</td>
<td></td>
<td>0.644</td>
<td>1.426</td>
</tr>
<tr>
<td>I2</td>
<td>0.626</td>
<td>0.035</td>
<td>0.567</td>
<td>1.254</td>
</tr>
<tr>
<td>I3</td>
<td>1.021</td>
<td>0.045</td>
<td>0.690</td>
<td>1.414</td>
</tr>
<tr>
<td>I4</td>
<td>1.002</td>
<td>0.053</td>
<td>0.444</td>
<td>1.947</td>
</tr>
<tr>
<td>I5</td>
<td>1.100</td>
<td>0.047</td>
<td>0.770</td>
<td>1.373</td>
</tr>
<tr>
<td>I6</td>
<td>1.297</td>
<td>0.052</td>
<td>0.720</td>
<td>1.632</td>
</tr>
<tr>
<td>I7</td>
<td>0.720</td>
<td>0.041</td>
<td>0.445</td>
<td>1.430</td>
</tr>
</tbody>
</table>

*Figure 20. Scalar model of the SIBS for Asian American students.*
Table 32: Parameter Estimates for Scalar Invariance on the SIBS for Latino Hispanic Students

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Unstandardized Loadings</th>
<th>Standard Error</th>
<th>Standardized Loadings</th>
<th>Unstandardized Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>I1</td>
<td>1.000</td>
<td>0.653</td>
<td>1.309</td>
<td></td>
</tr>
<tr>
<td>I2</td>
<td>0.626</td>
<td>0.035</td>
<td>0.576</td>
<td>1.166</td>
</tr>
<tr>
<td>I3</td>
<td>1.021</td>
<td>0.045</td>
<td>0.650</td>
<td>1.333</td>
</tr>
<tr>
<td>I4</td>
<td>1.002</td>
<td>0.053</td>
<td>0.385</td>
<td>2.063</td>
</tr>
<tr>
<td>I5</td>
<td>1.100</td>
<td>0.047</td>
<td>0.718</td>
<td>1.297</td>
</tr>
<tr>
<td>I6</td>
<td>1.297</td>
<td>0.052</td>
<td>0.765</td>
<td>1.395</td>
</tr>
<tr>
<td>I7</td>
<td>0.720</td>
<td>0.041</td>
<td>0.460</td>
<td>1.302</td>
</tr>
</tbody>
</table>

*Figure 21. Scalar model of the SIBS for Latino Hispanic students.*
Table 33

Parameter Estimates for Scalar Invariance on the SIBS for Multiracial Students

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Unstandardized Loadings</th>
<th>Standard Error</th>
<th>Standardized Loadings</th>
<th>Unstandardized Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>I1</td>
<td>1.000</td>
<td></td>
<td>0.657</td>
<td>1.200</td>
</tr>
<tr>
<td>I2</td>
<td>0.626</td>
<td>0.035</td>
<td>0.579</td>
<td>1.101</td>
</tr>
<tr>
<td>I3</td>
<td>1.021</td>
<td>0.045</td>
<td>0.599</td>
<td>1.238</td>
</tr>
<tr>
<td>I4</td>
<td>1.002</td>
<td>0.053</td>
<td>0.334</td>
<td>1.238</td>
</tr>
<tr>
<td>I5</td>
<td>1.100</td>
<td>0.047</td>
<td>0.698</td>
<td>1.738</td>
</tr>
<tr>
<td>I6</td>
<td>1.297</td>
<td>0.052</td>
<td>0.822</td>
<td>1.193</td>
</tr>
<tr>
<td>I7</td>
<td>0.720</td>
<td>0.041</td>
<td>0.513</td>
<td>1.242</td>
</tr>
</tbody>
</table>

Figure 22. Scalar model of the SIBS for multiracial students.
Table 34
Parameter Estimates for Scalar Invariance on the SIBS for European American Students

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Unstandardized Loadings</th>
<th>Standard Error</th>
<th>Standardized Loadings</th>
<th>Unstandardized Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>I1</td>
<td>1.000</td>
<td>0.000</td>
<td>0.645</td>
<td>1.368</td>
</tr>
<tr>
<td>I2</td>
<td>0.626</td>
<td>0.035</td>
<td>0.533</td>
<td>1.227</td>
</tr>
<tr>
<td>I3</td>
<td>1.021</td>
<td>0.045</td>
<td>0.639</td>
<td>1.396</td>
</tr>
<tr>
<td>I4</td>
<td>1.002</td>
<td>0.053</td>
<td>0.417</td>
<td>2.172</td>
</tr>
<tr>
<td>I5</td>
<td>1.100</td>
<td>0.047</td>
<td>0.681</td>
<td>1.418</td>
</tr>
<tr>
<td>I6</td>
<td>1.297</td>
<td>0.052</td>
<td>0.686</td>
<td>1.680</td>
</tr>
<tr>
<td>I7</td>
<td>0.720</td>
<td>0.041</td>
<td>0.469</td>
<td>1.375</td>
</tr>
</tbody>
</table>

*Figure 23. Scalar model of the SIBS for European American students.*