2016-12-01

Evidence for the Validity of the Student Risk Screening Scale in Middle School: A Multilevel Confirmatory Factor Analysis

Matthew Porter Wilcox
Brigham Young University

Evidence for the Validity of the Student Risk Screening Scale in Middle School:

A Multilevel Confirmatory Factor Analysis

Matthew Porter Wilcox

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Richard R. Sudweeks, Chair
Lane Fischer
Michael J. Richardson
Ellie Young
Joseph A. Olsen

Educational Inquiry, Measurement, and Evaluation

Brigham Young University

ABSTRACT

Evidence for the Validity of the Student Risk Screening Scale in Middle School: A Multilevel Confirmatory Factor Analysis

Matthew Porter Wilcox
Educational Inquiry, Measurement, and Evaluation, BYU
Doctor of Philosophy

The Student Risk Screening Scale—Internalizing/Externalizing (SRSS-IE) was developed to screen elementary-aged students for Emotional and Behavioral Disorders (EBD). Its use has been extended to middle schools with little evidence that it measures the same constructs as in elementary schools. Scores of a middle school population from the SRSS-IE are analyzed with Multilevel Confirmatory Factor Analysis (MCFA) to examine its factor structure, factorial invariance between females and males, and its reliability. Several MCFA models are specified, and compared, with two retained for further analysis. The first model is a single-level model with chi-square and standard errors adjusted for the clustered nature of the data. The second model is a two-level model. Both support the hypothesized structure found in elementary populations of two factors (Externalizing and Internalizing). All items load on only one factor except Peer Rejection, which loads on both. Reliability is estimated for both models using several methods, which result in reliability coefficients ranging between .89-.98. Both models also show evidence of Configural, Metric, and Scalar invariance between females and males. While more research is needed to provide other kinds of evidence of validity in middle school populations, results from this study indicate that the SRSS-IE is an effective screening tool for EBD.

Keywords: Student Risk Screening Scale, universal screening, emotional and behavioral disorders, reliability, validity, multilevel modeling, structural equation modeling

ACKNOWLEDGEMENTS

Thanks and credit go to the members of my committee, the school district that was gracious enough to allow us to do research there, and my lovely bride. What more can I say?

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Confirmatory Factor Analysis (CFA)

Conner Rating Scale (CRS)

Development and Well-being Assessment (DAWBA)

Differential Item Functioning (DIF)

Emotional and Behavioral Disorders (EBD)

Exploratory Factor Analysis (EFA)

Factor Mixture Modeling (FMM)

Item Response Theory (IRT)

Maximum Likelihood (ML)

Measurement Invariance/Equivalence (MIE)

Multilevel Structural Equation Modeling (MSEM)

Multiple Indicators Multiple Causes (MIMIC)

Múthen's Maximum Likelihood (MUML)

National Council on Measurement in Education (NCME)

Office Discipline Referral (ODR)

Robust Maximum Likelihood (MLR)

Structural Equation Modeling (SEM)

Student Difficulties Questionnaire (SDQ)

Student Risk Screening Scale (SRSS)

Student Risk Screening Scale--Internalizing/Externalizing (SRSS_IE12)

Student Scale for Behavioral Disorders (SSBD)

CHAPTER 1: INTRODUCTION

Over the past 15 years, several national and state education initiatives such as No Child Left Behind (NCLB) and the Common Core State Standards have attempted to raise student achievement by focusing primarily on teacher, administrator, and school accountability in providing quality instruction. However, these approaches alone have generally ignored the connection between a student's *social and emotional* development and their *academic* performance (Elias & Arnold, 2006). An increasing body of research shows that providing timely school-wide social and emotional instruction is an effective component in helping students meet academic goals, as well as improve their quality of life outside of school (Durlak & Weissberg, 2007; Hoffman, 2009; Zins, Weissberg, Wang, & Walberg, 2004).

Background of the Problem

While school-wide social and emotional instruction can benefit all students, those who have developed or are at risk for Emotional and Behavioral Disorders (EBD) may need extra support in the form of class-specific and individual interventions. These disorders are a subset of mental health issues and are identified as

> (A) An inability to learn that cannot be explained by intellectual, sensory, or health factors. (B) An inability to build or maintain satisfactory interpersonal relationships with peers and teachers. (C) Inappropriate types of behavior or feelings under normal circumstances. (D) A general pervasive mood of unhappiness or depression. (E) A tendency to develop physical symptoms or fears associated with personal or school problems. (Code of Federal Regulations, 2012, Title 34, Section 300.7(c)(4)(i))

EBD can manifest in the form of externalizing, internalizing, or comorbid behaviors. Externalizing disorders tend to be more noticeable, and are generally exhibited through anti-social and aggressive behaviors (Stouthamer-Loeber & Loeber, 2002). Internalizing behaviors are usually harder to observe in classrooms and are most commonly manifest through anxiety and depression (Morris, Shah, & Morris, 2002). EBD can also be comorbid, meaning that both externalizing and internalizing behaviors are present in the same student (Kovacs & Devlin, 1998; Ollendick & King, 1994).

The presence of Emotional and Behavioral Disorders (EBD) in children has been shown to have substantial adverse effects on K-12 learning outcomes, and have been linked to significantly lower school performance, increased referrals for discipline issues, and higher dropout rates. Poor school performance, however, is only one of many negative effects. Outside of school, EBD is predictive of increased risk of other mental health issues, abuse, criminal behavior, and unemployment (Bullis & Yovanoff, 2006; Landrum, Tankersley, & Kaufman, 2003). Although such outcomes for those who have or are at-risk for developing EBD appear bleak, timely class-wide or individual interventions increase academic performance and greater social integration and employment prospects (Allen-DeBoer, Malmgren, & Glass, 2006). Further, the system for delivering these interventions is already in place. Although an estimated one-quarter of all students experience mental health issues at some point during K-12 (Egger & Angold, 2006), public schools already provide approximately 70-80% of mental health services to children and youth (Rones & Hoagwood, 2000). Thus, if schools could accurately identify students who have or are at-risk for EBD, they already have the framework for providing the interventions that can raise student achievement and increase the chances of lifetime success.

While the framework for assisting students with EBD may already be in place, there is not always a robust, systematic way to identify them. Too often, schools rely on an outdated discrepancy model, where a teacher refers a student to the school psychologist because of severe emotional or behavioral issues. There are several problems with this approach. This method of referral generally identifies those whose behavioral problems are obvious, leaving behind those who may be at-risk. Further, when a student’s behavior is allowed to worsen until it is detrimental enough to warrant a referral, they are generally less responsive to interventions than if they had been identified at an earlier stage.

A more efficient way to identify students who have or are at-risk of EBD is to proactively screen all students using a psychometrically validated teacher report form; this is referred to as universal screening. Each teacher fills out a screening instrument for each child in their class based on their observation of that student throughout the regular course of their time at school. A summed score is calculated, and those students whose scores fall above a predetermined cut score are referred to the school psychologist for observation, with the possibility of a diagnosis and accompanying intervention. This process allows each child to be considered for services; further, it quickly and inexpensively narrows the pool of potential students with issues without requiring the school psychologist to observe every child (Glover & Albers, 2007).

Glover and Albers (2007) provided practical guidelines for educators in evaluating universal screening assessments for use in their particular school. One of their conclusions is that

> Although significant advances in screening have led to improvements in the ability to identify and serve students, additional research is warranted to ensure that screening assessments are contextually relevant, psychometrically adequate, and usable. Current approaches are promising, but warrant further development and research. It is expected that much-needed future investigations will make an impact on universal screening policy and practice. (p. 128)

Thus, the appropriateness of the screening instrument is key to both the policy and practice of identifying and helping students in need of interventions. In other words, scores obtained from universal screening instruments should have sufficient evidence of reliability and validity (AERA et. al, 2014) to ensure that students are accurately identified.

There are several screening instruments commonly used in elementary schools by teachers to determine at-risk behaviors. These include the Student Risk Screening Scale (SRSS), the Student Screener for Behavioral Disorders (SSBD), the Strengths and Difficulties Questionnaire (SDQ), and the Behavioral Assessment for School Children (BASC). While the use of these screening tools promises the early identification of at-risk behaviors for elementary-aged students, they are increasingly being used to identify at-risk behaviors with middle and even high-school aged students. Because they were developed primarily for younger children, evidence of validity with middle schools populations is weak, incomplete, or even non-existent. Without strong evidence of reliability and validity with older ages, the scores from a screening instrument may systematically be interpreted as a false positive or false negative, resulting in wasted resources, and children in need of help not receiving it. Such an outcome in universal screening would be little better than the discrepancy model.

Problem Statement

Secondary schools may be using screening instruments for EBD that have not been adequately tested for an adolescent population, resulting in students who have developed or are at-risk for developing EBD potentially not receiving the interventions they need to be successful in school and life.

Purpose

The goal of this study was to examine the evidence for the reliability and validity of a universal screening instrument for Emotional and Behavioral Disorders with a middle school population.

Theoretical Framework

To evaluate the strength of the evidence for validity concerning the screeners above, criteria for establishing validity as found in the *Standards for Educational and Psychological Testing* (AERA et. al, 2014) are used. Sponsored by the American Educational Research Association (AERA), this work was written by a panel of experts from the National Council on Measurement in Education (NCME) and the American Psychological Association (APA) and is considered the standard for developing and using tests in education.

Validity. The *Standards* defines validity as “the degree to which evidence and theory support the interpretation of test scores” (AERA et al., 2014, p. 9). For example, does a math test measure a student's math ability, or some other trait like literacy, cognitive speed, test-wiseness, and so forth. Providing evidence for this aspect of testing is a question of validity. In earlier versions of the *Standards*, validity was thought of as including different types or facets, such as content-related validity, criterion-related validity, and construct-related validity. The current edition of the *Standards*, however, states that validity is unitary, with five different kinds of evidence to support the claim of validity. The five types of evidence include evidence based on (a) test content, (b) response processes, (c) internal structure, (d) relations to other variables, (e) and consequences of testing. Evidence of validity based on relations to other variables is further subdivided into predictive, concurrent, convergent, and discriminant evidence. Because validity is a construct that cannot be directly observed or proven a researcher builds a case for validity by providing evidence from these five sources (AERA et al., 2014).

Reliability. Reliability is an important part of establishing evidence for the validity of an instrument. The *Standards* define reliability as the "consistency of such measurements when the testing procedure is repeated" (p. 25). In other words, if the instrument were repeatedly given under similar testing circumstances, it would produce similar results. While the *Standards* do not specify a method for determining reliability, reliability coefficients generally range between 0 and 1; lower coefficients indicate lower reliability with values above .80 indicating high reliability (Nunnally & Bernstein, 1994). Test-retest reliability, or computing the correlation between scores or ratings from two or more testing occasions for the same population on the same test, is an important indicator of reliability. Coefficients for test-retest reliability tend to be lower than those reported for internal consistency because of measurement error and changes in the students as they respond to interventions.

A case for the validity of the interpretation of the scores for a screening instrument is not firmly established with a single study but is built over multiple studies. As such, this study will find areas where there is a paucity of research concerning the evidence for the validity of one of the four screening instruments listed earlier. The following review of the literature will examine the evidence for the reliability and validity of these four screening instruments according to the criteria found in the *Standards for Educational and Psychological Testing*. The justification for retaining the SRSS for further study over the other three screeners is provided.

Summary

While providing school-wide social and emotional support can increase student achievement, those who have or are at-risk for EBD may need additional class-wide or individual interventions. Identifying these students is possible through universal screening of all students using a teacher-report form that is psychometrically sound.

CHAPTER 2: REVIEW OF THE LITERATURE

As mentioned above, reliable and valid measures are needed to correctly identify children who have or are at-risk for Emotional and Behavioral Disorders (EBD). To this end, the literature has been reviewed concerning research on the reliability and validity of four primary screening instruments of EBD. The SRSS is retained for further study concerning evidence for its internal structure as found in the *Standards for Educational and Psychological Testing*.

Search Method

Two different searches were conducted to find relevant literature concerning screening tools for Emotional and Behavioral Disorders (EBD). For the first search, ERIC and PsychINFO Databases, Google Scholar, and Electronic Theses and Dissertations were employed. This search was conducted by using the names of the four most common universal screening instruments for Emotional and Behavioral Disorders (EBD) as keywords: *Behavioral Assessment System for Children* (BASC), *Systematic Screening for Behavioral Disorders* (SSBD), *Strengths and Difficulties Questionnaire* (SDQ), and the *Student Risk Screening Scale* (SRSS). This search returned a total of 1,436 results. The terms *test reliability* and *test validity* were then added to the search, which narrowed the results to 236.

The second search was conducted to find other relevant articles outside of these four primary screening tools. Using the thesaurus function in ERIC and PsychINFO, the following terms were included: *emotional disorders, behavioral problems, screening tests, elementary schools, secondary schools, at-risk persons, social development, test* *validity,* and *test reliability*. The search returned 48 articles, for a total of 284 articles, from both searches for consideration in this study.

From these 284 articles, publications were chosen for inclusion only if the primary focus of the article was the validation of the instrument. Indicators of test validity included keywords such as *construct, convergent, divergent, or discriminant validity; reliability*; *exploratory*, *confirmatory factor analysis, or factor structure;* or *Item Response Theory*. Based on these criteria, 51 articles are retained for inclusion in the following section.

Results

After having narrowed the pool of pubilcations for review, the publications for inclusion in this review were categorized by screening instrument. The section for each instrument contains a brief description of the screener itself, a review of the relevant articles in regards to evidence of reliability and validity, and an analysis of the strength of said evidence as compared to the standards.

Student Screener for Behavioral Disorders (SSBD). The SSBD was developed for elementary-aged populations by Walker and Severson (1992) and is considered state of the art in screening for EBD (Gresham, Lane, & Lambros, 2000; Kauffman, 2001). It has three gates through which students pass to identify those most in need of services. In the first stage, the teacher considers all students and ranks the top ten students who exhibit externalizing behaviors and the top ten who exhibit internalizing behaviors (students can be placed on both lists). From this initial ranking, the top three from each list pass to the second gate, or round, of assessment. In the second gate, the teacher fills out a 33-item *Critical Events Checklist* and a 23-item *Combined Frequency Index* for each of the three internalizing and externalizing students. Scores for students which exceed the cutoff move to the third stage (Walker & Severson, 1992). The final stage involves observation by the school psychologist and possible intervention.

Multiple studies have provided evidence supporting the use of the SSBD in elementary settings. For example, Gresham et al.(2000), and Kauffman (2001) found that the SSBD was effective at predicting both externalizing and internalizing behaviors in elementary students. Each study showed reliability estimates for the screeners to be high ($\alpha > .90$). Even though it is considered the gold standard in assessing EBD, the SSBD has some practical limitations. For example, it may not be as feasible as some of the other instruments because it takes teachers more than an hour to complete(Lane, Robertson, Kalberg, Lambert, Crnobori, & Bruhn, 2009). Further, while the SSBD does allow consideration of all students for services, only six total students are considered for interventions. This may leave other students who are at-risk from receiving needed services (Lane et al., 2008).

Although the screener was originally developed for elementary-aged populations, others have examined the possibility of its use in middle or junior high schools. The first of these studies examined the reliability, and convergent and discriminant validity of stage two scores in a secondary setting (Caldarella, Young, Richardson, Young, & Young, 2008). The researchers gathered data from 2,146 students attending middle and junior high schools from grades six to nine in rural and suburban Utah. Caldarella et al. found moderate to strong evidence for the reliability of the stage two screeners ($\alpha =.71$). They also found that the 123 students who were moved to stage two screening had more ODRs and lower GPAs than students who were not nominated. They further correlated the scores from the stage two screeners with the Social Skills Rating System (SSRS; not to be confused with the Student Risk Screening Scale, or SRSS) and found moderate to high correlations, indicating evidence based on relationships to other variables.

Richardson, Caldarella, Young, Young, and Young (2009) replicated the research from Caldarella et. al (2008) while extending the research question to deal not only with the convergent validity of the SSBD, but also the discriminant validity with a predominantly Caucasian suburban middle and junior high schools in the Intermountain West. They found significant correlations between nomination for EBD, higher ODR, and lower GPA, as well as evidence of convergent validity through comparison of scores to the SSRS. To explore discriminant validity, the authors compared the stage two results of the SSBD to the Achenbach System of Empirically Based Assessment (AESBA) and found small correlations between the emotional and behavioral subscales. The results indicated that in secondary-aged settings, the SSBD measures similar constructs to that of the Social Skills Rating Scale (SSRS) and different constructs than the AESBA. As in the previous study, this evidence based on relations to other variables that the SSBD measures a similar construct as the SSRS.

Young, Sabbah, Young, Reiser, and Richardson (2010) examined the reliability of the SSBD stage one and stage two scores in a secondary setting across gender. They further explored the interaction of gender and type of EBD (internalizing and externalizing). Over the span of three years, they collected over 15,000 scores from five different schools grade 6 to 9 throughout the rural United States, the majority of which is Caucasian. While there were no significant differences in reliability of the screener regarding gender, the authors found that males were three times more likely to be nominated for at-risk behaviors than females. They point out that the cause for having so many more males nominated for at-risk behaviors is unknown, but acknowledge that the SSBD itself may not capture female at-risk behaviors. They admit that the results do not necessarily point to a differential functioning of the instrument across gender; the difference may be due to the divide along externalizing/internalizing lines rather than male/female. This means that because males are more likely to be externalizers, they were more liable to be identified for at-risk behaviors.

While this research contributes valuable knowledge of how the screener functions in middle and junior high schools, these final three studies only provide evidence based on relationships to other variables; thus, other lines of evidence are needed in order to provide a stronger case for the validity of the SSBD in junior and middle schools.

Strengths and Difficulties Questionnaire (SDQ). The Strengths and Difficulties Questionnaire (SDQ) was developed in Great Britain by Goodman, (1997, 1999) to address specific problems inherent in surveys. Number of items, reading level, and the negative feelings that can stem from items dealing with EBD were addressed by limiting the number of items to 25, writing the items on a fifth-grade reading level, and including items regarding positive behavior. Unlike the SSBD, which was initially created to be used in elementary schools, the SDQ was designed to identify students between the ages of 5 and 17. Further, the SDQ has parallel forms for the parent and teacher, and if desired, student self-report forms for those aged 11 to 17. The 25 items from the SDQ are divided into five subscales: (a) Prosocial behavior, (b) Hyper-activity-Inattention, (c) Emotional Symptoms, (d) Conduct Problems, and (e) Peer Relationship Problems. While any number of students can be identified for at-risk behaviors, the normative sample estimates that 80% receive scores considered normal, 10% considered at-risk, and 10% abnormal or severe. Since its initial development, the SDQ has been translated into over 60 different languages and is used widely throughout the world.

Due to a number of languages into which the SDQ has been rendered, research concerning evidence for its reliability and validity has been conducted all over the globe, and investigations concerning its reliability and validity are more numerous than any of the other screeners. An overview of the literature reveals that evidence based on internal structure, specifically concerning the SDQ's factor structure, has dominated the literature. Hypothetically, the SDQ has five constructs, with each construct being measured by one of the five subscales. Goodman (2001), Smedje, Broman, Hetta, and von Knorring (1999), and Thabet, Stretch, and Vostanis (2000) found evidence for a five-factor structure through Explanatory Factor Analysis (EFA) in British, Swedish, and Arabian populations, respectively. However, Dickey & Blumberg (2004), in the first study examining the SDQ in United States elementary and secondary populations, used Confirmatory Factor Analysis (CFA) instead of Explanatory Factor Analysis (EFA) to establish the structure mentioned above and found that a three-factor model (externalizing, internalizing, and prosocial) was a better fit than the five-factor models proposed. Initially the authors hypothesized that the differences in factor structures conducted in Europe with EFA and the results in the United States with CFA were due to cultural differences. However, Rønning, Handegaard, Sourander, and Mørch (2004) found the three-factor structure a better fit with an Australian population using CFA. Also, Van Leeuwen, Meerschaert, Bosmans, De Medts, and Braet (2006) in a study of the

SDQ in a Russian population found an acceptable overall fit, but consistently low factor loadings.

Overall, forming concrete conclusions about the correct factor structure based on the literature is complicated; researchers have chosen to examine only one of the three forms, or any combination of the three. It is further complicated by analysis (EFA, CFA, or both) and population (language, ethnicity, race, age group, etc.). Due to these factors, results are mixed. For example, Matsuishi et al. (2008) found a good fit for the five-factor model with scores from the self-report form based on scores from 2899 Japanese children aged four to twelve. Because these researchers used Principal Components Analysis (similar to EFA) to conduct their analysis, a confirmatory analysis is needed to support their findings. Yao et al. (2009) also examined scores of the student report form obtained from 1,135 adolescents aged 13 to 17 in Mainland China using CFA and found good fit for the five-factor model. Van Roy, Veenstra, and Clench-Aas (2008) examined the scores obtained from the student report form with 26,369, Norwegian pre-early and late adolescents 10-19 years of age. Although they found modest fit with the five-factor model through CFA, they recommend improvements be made to improve scale properties including internal consistency. One shortcoming of these studies, however, is that the researchers generally only examined one model; they may have found better fit by examining and comparing multiple alternate models.

Several other studies have been conducted on the student the self-report form in different populations that support the three-factor model. For example, Di Riso et al. (2010) provided evidence for the three-factor model through CFA with scores of 1394 Italian children aged 8 to 10 on the student form. Ruchkin, Jones, Vermeiren, and

Schwab-Stone (2008) reporting on scores from among 4,761 American 6th through 10th graders also found weak loadings and reliability estimates of the student form with the five-factor model. They found the best fit with a three-factor model through a CFA approach. Finally, Haynes, Gilmore, Shochet, Campbell, and Roberts (2013) used both EFA and CFA to assess the fit of the three and five-factor models on the self-report form among 128 Australian children between the ages of nine and fourteen diagnosed with an intellectual disability. Although their analysis provided evidence that the three-factor model had better fit, several aspects of their study are questionable. First, the small sample size and sample type make validity generalization flimsy. Second, although conducting both EFA and CFA can provide robust evidence, the general practice is to obtain a sample large enough to split in half, with one-half used for EFA and the other for CFA. To conduct both analyses on the full sample is more likely to capitalize on chance and increase the probability that the CFA will confirm the results of the EFA. The researchers make no indication of splitting the sample.

Several researchers also examined the factor structures of the teacher and parent forms. For example, Van Roy et al. (2008) in their study of the SDQ among children in Norway were also able to collect 6,645 parent and teacher forms out of the 26,369 youth they tested. Although they found modest fit with the five-factor structure for the student form, results of their analysis on the parent and teacher forms were inconclusive. Ezpeleta, Granero, la Osa, Penelo, and Domènech (2013) collected parent and teacher forms for 1341 Spanish three-year-olds. Through CFA they found acceptable fit for the five-factor model, as well as good fit for the three-factor model.

Hill and Hughes, (2007) examined the scores of the SDQ for a racially diverse population in Texas. They conducted a CFA on 784 results from the parent and teacher forms and found marginal fit for both. However, they noted that some items loaded onto different factors depending on the form, and urged further development of the instrument. Ruchkin, Koposov, Vermeiren, and Schwab-Stone (2012) examined the structure of the teacher and student forms with a Russian population of 528 children between grades 6 and 10. Using CFA, they found that both three and five-factor model had good fit. Finally, McCrory and Layte (2012) compared the fit of four different models based on the results of the parent form for 8,514 nine-year-olds in Ireland. Using CFA they found that the basic five-factor structure originally proposed by Goodman (2001) as best fitting, with the three-factor structure suggested by Dickey and Blumberg (2004) the worst.

The research presented as to whether the SDQ has either three or five factors is and may continue to be inconclusive. Perhaps sensing this, Goodman, Lamping, and Ploubidis (2010) claimed that both models might be more appropriate depending on the context. The three-factor model would be more appropriate when dealing with populations where EBD was normally distributed, and the five-factor model may be appropriate when studying known high-risk populations. Evidence supporting the number and name of the different constructs is necessary as researchers attempt to provide evidence for the validity of the instrument. For example, several of the previous studies found only modest support for convergent evidence and poor support for discriminate evidence (Ezpeleta et al., 2013; Hill & Hughes, 2007; Yao et al., 2009). This may be due to the way they are grouping the items to form constructs based on their analysis of the factor structure.

Hagquist (2007) also found problems with the scale when he analyzed the scores on the self-report form of 8,838 12 to 18-year-old Swedish youth. He examined each of the five subscales using the Rasch model and concluded that the scale is in need of further improvement due to multiple misfitting items. However, if the constructs are identified incorrectly and the items inappropriately grouped, the results may be inaccurate. It may be that the scale does need further revision, but Hagquist's results may also be due to misspecified item groupings. For this reason, the issue of constructs continues to be an important topic for research.

Lastly, three publications focused on evidence for validity other than internal structure. Jee et al. (2011) explored the possibility of using the SDQ to identify children in foster care for EBD. Over the course of two years, they had 212 foster children between the ages of 11 and 17 and their foster parents fill out the respective forms from the SDQ, and found that the detection rate of at-risk youth doubled. Goodman and Goodman (2009) obtained scores from teacher, parent, and student forms for 7,483 British youth aged 11 to 16; after three years, clinicians reassessed these students for EBD. Those identified by the SDQ as at risk or high risk had a higher rate of psychopathology as judged by the clinical diagnosis. Goodman and Goodman (2011) correlated the scores from the self-report form of 18,425 youth ages 5 to 16 with diagnostic interviews from the Development and Well-being Assessment (DAWBA). Over the course of three years, they found that the scores obtained on the SDQ predicted scores on the DAWBA between 89 and 90 percent.

The literature provides substantial evidence concerning the SDQ's internal consistency, moderate evidence based on relationships to other variables, and inconclusive evidence based on internal structure. However, because these studies were conducted with various populations, spanning diverse age groups and ethnicities, or using different combinations of the various forms, more work needs to be done to clarify the research already completed. Further, due to the range of results from various researchers, evidence for the internal structure of the SDQ in its current form cannot be established across all populations. Research is also needed to provide other kinds of evidence of validity.

Behavior Assessment for Children, Second Edition: Behavioral and Emotional Screening System (BASC-2 BESS). The BASC Second Edition (Reynolds & Kamphaus, 2004) contains over 400 items and is the source from which several shorter scales have been derived, including the Behavioral and Emotional Screening System (BESS; Kamphaus & Reynolds, 2007). Like the SDQ, the BESS has teacher, parent, and student forms that range in length from 25 to 30 items with subscales intended to measure internalizing problems, externalizing problems, school problems, adaptive skills/personal adjustment, and inattention/hyperactivity. Teacher and parent forms have pre-K and K-12 versions, while the student form has only one version for grades three to twelve. The BESS has also been made more accessible to those who have difficulty reading through audio recordings of the items and has been translated into Spanish.

Kamphaus and Reynolds (2007) used a nationally representative sample of 5,888 children, ages three to seventeen years, to show initial evidence for reliability and validity. They have demonstrated that the BESS has strong internal consistency and test-retest reliability. However, the researchers fail to explain why they use split-half reliability to determine internal consistency. This method splits the sample and then correlates the scores from the two halves, which can produce erratic results depending on the manner in which the sample was split. Further, the scores indicate convergent and discriminant evidence when compared to Achenbach's Empirically Based Assessment Child Behavior Checklist (ASEBA) and Conner Rating Scale—Revised (Conners, 1997).

Dowdy et al. (2011) add to the evidence of validity for the BESS by using the normative sample from Kamphaus and Reynolds (2007) investigate the factor structure. Dowdy et al. split the sample and conducted an EFA and CFA on the two halves of the results from the parent forms and found good fit for a four-factor structure. Dever, Mays, Kamphaus, and Dowdy (2012) examined the factor structure of the teacher form based on the scores from a nationally representative sample of 2,582 students aged six to twelve, and offer evidence for a four-factor model. Dowdy, Chin, Twyford, and Dever (2011) also found a four-factor solution for the student form but obtained the result not only from a nationally represented sample but through a second CFA on results from 273 predominantly Hispanic students ages seven through twelve. These studies provide substantial evidence for validity based on internal structure.

Other researchers have examined the evidence based on relationships to other variables. For example, Renshaw et al. (2009) found that a negative correlation with the academic, behavioral and engagement marks of 48 third and fourth graders in California. Although this research lacks validity generalizability due to the low number of participants, others have found similar results with larger samples. For example, Kamphaus, Distefano, Dowdy, Eklund, and Dunn (2010) reported that the higher scores of 472 elementary students from Los Angeles on the BESS teacher form were related to lower GPA and academic achievement test scores. They further reported high internal reliability (α=.96). King, Reschly, and Appleton (2012) also found high internal consistency for the parent, student, and teacher forms with a sample of 496 elementary students from the Southeast. However, the correlations between the scores of the three forms were weak. They further found that the forms did not always correlate with the same variables. For example, high BESS scores from student forms correlated with more ODRs, and lower attendance and reading ability, while parent forms correlated with attendance and reading, but not ODRs.

Other researchers focused their efforts on populations other than elementary students. Chin, Dowdy, and Quirk (2012) examined the results of the teacher and student forms from 694 sixth and seventh graders and found high internal consistency for both (α=.82; student, α=.92; teacher). Kamphaus et al. (2010) further confirmed Kamphaus and Reynolds (2007) evidence for convergent and discriminant evidence between the BESS and ASEBA and the TRS-R. Dowdy, Chin, and Quirk (2013) have provided initial proof of validity based on its relationship to other variables in preschool. They compared the scores of the BESS teacher report form of 65 predominantly Latino three-year-olds with their scores on the Ages and Stages Questionnaire (ASQ-SE; Squires, Bricker, & Twombly, 2003) and the Peabody Picture Vocabulary Test--Fourth Edition (Dunn & Dunn, 2007). The researchers found a significant link between high BESS scores and lower levels of school readiness and vocabulary. However, due to the low number of participants, these results need to be replicated with a larger sample. Finally, Dowdy, Dever, Distefano, and Chin (2011) compared the scores of the BESS teacher report form of 142 native English speaking and 110 limited English Proficiency (LEP) students from

Kindergarten to fifth grade and found no evidence of differential item functioning. This provides evidence that the results from the screener are not biased towards LEP students.

Research concerning the BESS has consistently reported high values for internal consistency and test-retest reliability. Further, there is reasonable evidence to support the claim of validity based on internal structure and relationship to other variables. However, upon closer examination, the populations from which scores are derived have either come from the normative sample originally used in by Kamphaus and Reynolds (2007) or have been drawn primarily from predominantly Latino populations in California. This limits the generalizability of the use of the screener to different populations.

Student Risk Screening Scale (SRSS). The SRSS was developed by Drummond (1994) and consists of only seven items, and unlike the other screeners, it is openly licensed and free for use by anyone. Further, contrary to the SSBD, which only permits six students to pass to the second gate, the SRSS has the potential to recommend as many students that are at risk or who have developed EBD. However, a significant disadvantage of the SRSS is that six of the seven items address externalizing behaviors. Only one item addresses internalizing behaviors. This implies that the screener may not be as efficient as identifying those with internalizing disorders.

Although originally developed for use in elementary schools, several articles have been published providing evidence of validity in secondary school settings. For example Lane, Parks, Kalberg, and Carter (2007) examined the internal consistency, test-retest reliability, and convergent validity of the Student Risk Screening Scale (SRSS) with a sample of 500 predominantly Caucasian middle school students in a rural area of Tennessee. Lane et al. further examined the correlation between the scores from the

SRSS with the number of Office Discipline Referrals (ODR) and student Grade Point Average (GPA) with 528 students in a diverse urban middle school in the same state. The scores obtained from the rural middle school on the SRSS showed internal consistency levels above .70 as well as high test-retest reliability. The researchers also found high correlations between the SRSS and the externalizing subscales on Strengths and Difficulties Questionnaire (SDQ), and links between higher scores from the SRSS to a greater number of ODRs and lower GPA.

Lane, Bruhn, Eisner, and Robertson Kalberg (2010) further examined the validity of scores drawn from the SRSS in middle schools but focused solely on a diverse urban population. They tested 534 middle school students in grades 5 through 8 in Tennessee. As in the previous study, the authors found high internal consistency, moderate to high test-retest reliability, and significant correlations between the SRSS, ODR, and GPA. Lane, Robertson Kalberg, Parks, and Carter (2008) administered the SRSS to the teachers of 674 high school students from tenth to twelfth grades in Tennessee and found adequate convergent and discriminant evidence when compared to the various subscales of the SDQ. They also reported similar correlations between scores on the SRSS and ODRs and GPAs.

Related research comparing the scores of SRSS to different criterion and other measures were conducted in elementary settings. Ennis, Lane, and Oakes (2011) explored reliability and convergent evidence of the SRSS with a racially diverse elementary population of 448 Kindergarten through fourth graders in urban Tennessee. The results from the SRSS showed high internal consistency ($\alpha = .86$), and moderate to strong test-retest reliability, with α ranging between .60 and .78. Further, the scores from the SRSS were highly correlated with similar subscales from the SSBD and SDQ. Menzies and Lane (2011) correlated the scores of the SRSS from 286 K-6 students in California with Office Disciplinary Referrals, measures of self-control skills, and second measures of proficiency in language arts. The results indicate that higher scores on the SRSS are linked to a higher number of ODRs and lower ability in self-control and language arts. The scores of the SRSS from 1,142 elementary students attending a diverse urban Midwest school showed a moderate negative correlation with oral reading fluency scores (Oakes et al. (2010).

Two sets of researchers have explored the extent to which the SRSS can identify internalizing disorders in elementary populations. Lane et al. (2008) compared the SRSS to the second stage screeners of the SSBD from 73 K-2 teachers on 578 students from a racially diverse setting in Tennessee. There was no statistical difference between the screeners regarding nomination for externalizing behaviors. However, the SSBD was far more accurate in identifying internalizing behaviors. Lane et al. (2009) replicated this study with a larger sample over more grade levels. Over the course of a year, they administered the SRSS and SSBD several times to the teachers of 2,588 K-5 students in Tennessee. The results again indicate similar scores for identifying externalizing behaviors between the two screeners, with the scores from the SSBD more accurately identifying internalizing behaviors than the SRSS.

Because the SRSS lacks the ability to identify more students with internalizing behaviors accurately, Lane added seven items to address the shortcoming (Lane, Menzies, et al., 2012). Her research associates tested the improved scale with 2,460 elementary students from four racially diverse schools located in Arizona and California.

Based partially on the results of a principal component analysis, two items were removed from the scale. They examined the psychometric properties of the new improved screener, and correlations increased significantly between the scores of the internalizing items with corresponding subscales from the SSBD and SDQ. Lane et al. (2012) conducted a similar study with 2061 K-4 students in rural and urban districts in the Southern United States. As with Lane et al. (2012), the results of the principal components analysis led the researchers to retain five of the new items for the scale and named it the SRSS-Internalizing/Externalizing 12 items (SRSS-IE). This same method was repeated with a middle school population, again showing a two-factor model with Peer Rejection loading on both factors (Lane et al., 2013). While these studies offer an important initial foray into the factor structure of the SRSS, further research is needed to confirm the structure.

The evidence for validity for the SRSS and SRSS-IE as a whole indicates persuasive evidence for the reliability of the instrument and validity based on relationships to other variables, mainly through the use of convergent and discriminant evidence. Nearly all the studies concerning this tool use an almost identical design, with the only variance being the population regarding place and age. In most cases, researchers correlated the results of the SRSS-IE to ODRs, GPAs, and the results of another scale such as the SSBD and SRSS. While this approach does build a strong case for the use of the instrument on diverse populations, other lines of evidence could help to strengthen arguments for validity. Further, more research is needed to establish its validity in secondary settings, and more research is required in all settings to provide other types of evidence of validity.

Overall analysis. The strength of evidence for validity varies from screener to screener; however, there are also some overall trends and patterns across all the screeners. First, the majority of the published articles provided evidence of reliability through reporting internal and test-retest reliability coefficients. Although the literature universally shows moderate-high to high reliability estimates using Cronbach's coefficient α, these coefficients may be misleading. For α to accurately measure the internal consistency of any instrument, researchers must satisfy certain assumptions. All too often, the statistic, if performed with no investigation as to whether the assumptions have been met, results in a biased estimate. None of the studies that calculated α showed any sign that the assumptions for the statistic were explored or met.

Second, evidence of validity does not adequately represent the five areas set forth by the *Standards*. Only one study examined evidence based on test content (Lane et al., 2012), and only one study focused on consequences of testing (Young et al., 2010). Twenty-seven of the reviewed studies—the majority—provide evidence based on relationships to other variables. While 17 studies that explore evidence based on internal structure, the majority of these studies deal with the SDQ and are inconclusive in providing evidence for a generalizable factor structure across varying populations. The methodology used to determine internal structure poses another issue. Only one study used Item Response Theory (IRT) to examine the structure of one of these instruments (Hagquist, 2007).

Third, published research is not distributed across populations of varying ages. Twenty-seven studies used elementary populations, 20 of which focus solely on elementary aged students; the other seven studies include both elementary and secondary students. Only six studies have been done exclusively using secondary populations.

The emphasis on evidence based on relations to other variables and internal structure, especially in elementary populations, indicates that there are several areas of research regarding EBD screeners that could significantly advance their case for validity. Research focused primarily on evidence based on relationships to other variables; more research on evidence based on test content, response processes, internal structure, and the consequences of testing would add valuable information as to the validity of these screeners. Second, when reporting Cronbach's coefficient α, assumptions need to be investigated and reported. When the data do not support the assumptions, other means of estimating reliability should be used. Finally, while the number of studies conducted in secondary settings has increased in recent years, much more needs to be done to better understand how well these screeners function with junior high and high school students. In Fact Lane (2006) and Kalberg (2010) have made calls for further research in secondary aged populations.

Retained Screener and Research Questions

Based on the literature review, the Student Risk Screening Scale-Internalizing and Externalizing (SRSS-IE) was retained for the following reasons. First, unlike the BASC-2 and SSBD, the SRSS-IE is free. Education budgets almost always have more demands than resources; thus a free screener may be more likely to be used than one that costs money. Although the SDQ is also free, the SRSS has another advantage in that it is shorter than the other three scales, including the SDQ. Longer scales are an advantage from a strict measurement perspective; the more quality items a scale has, the greater the precision or reliability. However, measurement considerations must be weighed in the balance with the practical aspects of assessment. In the case of these screeners, teacher workload must be taken into account, and practitioners with little expendable time during the average day would be more likely to complete a screening tool that takes less time. Finally, the extensive work done by Kathleen Lane, as shown in the previous section in extending the scale to include internalizing items shows promise for greater identification of students who are at risk for EBD.

As noted earlier, evidence of validity in secondary settings is much needed, as well as more information as to how the scale functions differentially between males and females. This study examined the scores obtained from the SRSS-IE in a secondary setting, focusing primarily on evidence of internal structure. The following questions were addressed to investigate the evidence of internal structure.

To what extent do SRSS-IE scores from a middle-school population show evidence of:

1. A two-factor structure with Peer Rejection loading on both factors, as proposed by Lane et al. (2013)?
2. Internal consistency reliability?
3. Measurement equivalence/invariance (ME/I) between males and females?

Summary

A review of the literature concerning four screening instruments has been reported. Selected articles deal in evidence of validity for the screening instruments, and were evaluated against the criteria outlined in the *Standards for Educational and Psychological Testing.* Most research has been conducted with elementary-aged populations, and predominantly focuses on evidence of relationships to other variables. Cronbach's coefficient α was often reported as the estimate of internal consistency, but assumptions were not examined. The SRSS-IE is retained for further study.

CHAPTER 3: METHOD

The purpose of this research was to examine evidence of validity regarding the internal structure of the scores obtained from the Student Risk Screening Scale – Internalizing/Externalizing (SRSS-IE) in a middle school population. The following questions are addressed:

To what extent do SRSS-IE scores from a middle-school population show evidence of:

1. A two-factor structure with Peer Rejection loading on both factors, as proposed by Lane et al. (2013)?
2. Internal consistency reliability?
3. Measurement equivalence/invariance (ME/I) between males and females?

This section outlines the method of conducting the study to answer these questions, including the design, participants, measures, procedures, and data analysis.

Design

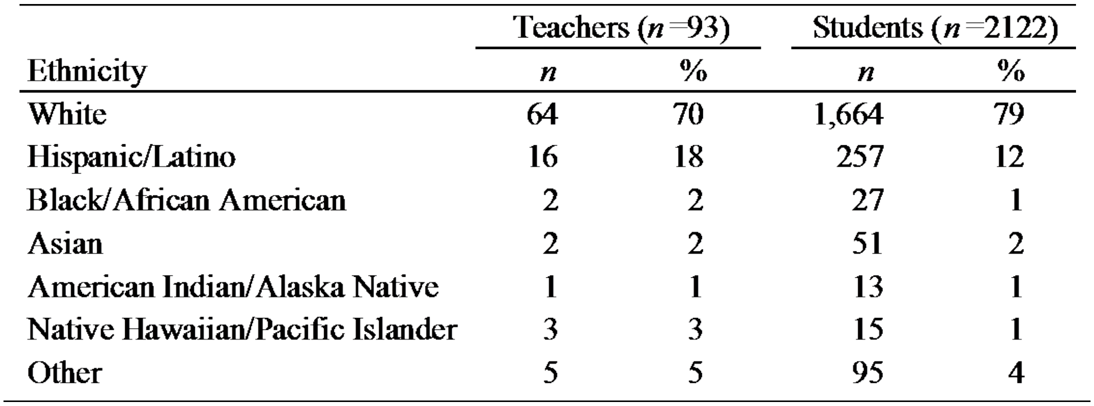
A cross-sectional design was used to obtain teachers' ratings of 2,122 middle school students from the SRSS-IE during the winter semester of the 2014-2015 school year. In general, cross-sectional designs allow researchers to quickly and inexpensively gather and analyze data concerning a particular population or subset of a population; this makes it an attractive, and in many instances necessary, means to provide evidence for more complex, in depth, or expensive studies.

Participants

Participants were 93 teachers (57 [61%] female; 36 [39%] male) from three middle schools (grades 6-8) in Utah. These teachers used the SRSS-IE to rate the observed behaviors of 2,122 students in their first-period class (1,042 [49 %] female, 1,080 [51%] male). As shown in Table 1, both teacher and student populations are predominantly white, with a notable Hispanic/Latino population.

Table 1

Teacher and Student Demographics



	Teachers (*n* =93)		Students (*n* =2122)	
Ethnicity	*n*	%	*n*	%
White	64	70	1,664	79
Hispanic/Latino	16	18	257	12
Black/African American	2	2	27	1
Asian	2	2	51	2
American Indian/Alaska Native	1	1	13	1
Native Hawaiian/Pacific Islander	3	3	15	1
Other	5	5	95	4

Measures

The *Student Risk Screening Scale* (formerly called the SRSS, now referred to as the SRSS-E7) was developed to identify elementary-aged students with conduct problems (Drummond, 1994). Teachers rate each student using a four-category scale based on the frequency they have observed the student engage in the specified behavior: 0 (*never*), 1 (*occasionally*), 2 (*sometimes*), 3 (*frequently*). Summed ratings have a potential range of 0 to 21; students receiving a rating of 9 or above are considered at-risk and are referred for further observation or assessment. While there is research that provides evidence of validity for this SRSS-E7, one criticism is that only one of its seven items directly addresses internalizing problems (Lane et al., 2007).

In 2012, researchers developed the *Student Risk Scale-Externalizing/Internalizing* (SRSS-IE14), which had the same basic items and rating scale as Drummond's original SRSS, but added seven new items to better screen for internalizing problems (Lane et al., 2013). Through the course of their research, Lane and associates reduced the 14 items to 12, resulting in the scale that is now referred to as the SRSS-IE. The 12 items on this scale include: (a) Stealing; (b) Lying, Cheating, Sneaking; (c) Behavior Problems; (d) Peer Rejection; (e) Low Academic Achievement; (e) Negative Attitude; (f) Aggressive Behavior; (g) Emotionally Flat; (h) Shy, Withdrawn; (i) Sad, Depressed; (j) Anxious; and (k) Lonely. Multiple studies have provided evidence internal consistency coefficients ranging from .76 to .87, and evidence of convergent validity with the Strengths and Difficulties Questionnaire (SDQ) and the Systematic Screening for Behavior Disorders (SSBD) (Ennis et al., 2011; Menzies & Lane, 2011; Oakes et al., 2010). Although initially designed for elementary-aged populations, similar results have been reported with scores obtained from secondary schools (Kalberg, Lane, Driscoll, & Wehby, 2010; Lane et al., 2010; Lane et al., 2007).

Procedures

Following approval by the Institutional Review Boards (IRB) for both the university and the participating school district, data were collected near the end of the 2014-2015 school year. In a single data-collecting session for each of the three participating schools, teachers rated their students' behaviors using the SRSS-IE. Researchers were present during the screening to ensure that all students were rated for each item; as a result, there is no missing data. To maintain the independence of observations, teachers only rated students from their first period. One pair of teachers each independently rated the same students for the first-period class that they team-taught. Including both sets of ratings for the same students would be a violation of the assumption of independence of observations for this study. To resolve this, one set of observations from these two teachers was randomly chosen and retained, and the other scores were removed from the analysis.

Data Analysis

As mentioned above, data obtained from the SRSS-IE are reported using four categories, but several items have very low counts. Table 3 shows the frequency counts and percentages of the response categories for each of the 12 items on the SRSS-IE.

For example, teachers reported having observed students frequently (the highest rating of 3) *Stealing* a total of 9 times in the entire sample of 2,122 students. Such a low observation rate for this item and category can have implications concerning the appropriateness of the number of response categories and estimation issues when trying to specify certain models. Such a low count may mean that having four categories for the item *Stealing* doesn't contribute to the understanding the latent trait, and that that three categories, rather than four, is more appropriate. However, reducing the number of categories will, to some extent, result in a loss of information. Further, low counts in a category can cause problems for model convergence, especially when examining Factorial Invariance where such a low count is divided among the groups. Linacre (2002) suggests that to avoid specification problems, categories with fewer than ten observations be combined with an adjacent category. Subsequently, for the *Stealing* item, all category 4 ratings were combined with category 3.

Table 2

Response Categories and Percentages

Item	Never	Rarely	Sometimes	Frequently
Steal	1962 *92.5%*	123 *5.8%*	28 *1.3%*	9 *0.4%*
Lie, Cheat, Sneak	1613 *76.0%*	267 *12.6%*	156 *7.4%*	86 *4.1%*
Behavior Problems	1627 *76.7%*	223 *10.5%*	171 *8.1%*	101 *4.8%*
Peer Rejection	1730 *81.5%*	245 *11.5%*	110 *5.2%*	37 *1.7%*
Low Academic Acheivment	1273 *60.0%*	342 *16.1%*	251 *11.8%*	256 *12.1%*
Negative Attitude	1614 *76.1%*	267 *12.6%*	145 *68.0%*	96 *4.5%*
Aggression Problems	1812 *85.4%*	195 *9.2%*	76 *3.6%*	39 *1.8%*
Emotionally Flat	1717 *80.9%*	221 *10.4%*	121 *5.7%*	63 *3.0%*
Shy, Withdrawn	1437 *66.7%*	300 *14.1%*	208 *9.8%*	177 *8.3%*
Sad, Depressed	1751 *82.5%*	222 *10.5%*	106 *5.0%*	43 *2.0%*
Anxious	1779 *83.8%*	202 *9.5%*	97 *4.6%*	44 *2.1%*
Lonely	1770 *83.4%*	206 *9.7%*	105 *4.9%*	41 *1.9%*
Averages	*1674* *79%*	*234* *11%*	*131* *6%*	*83* *4%*

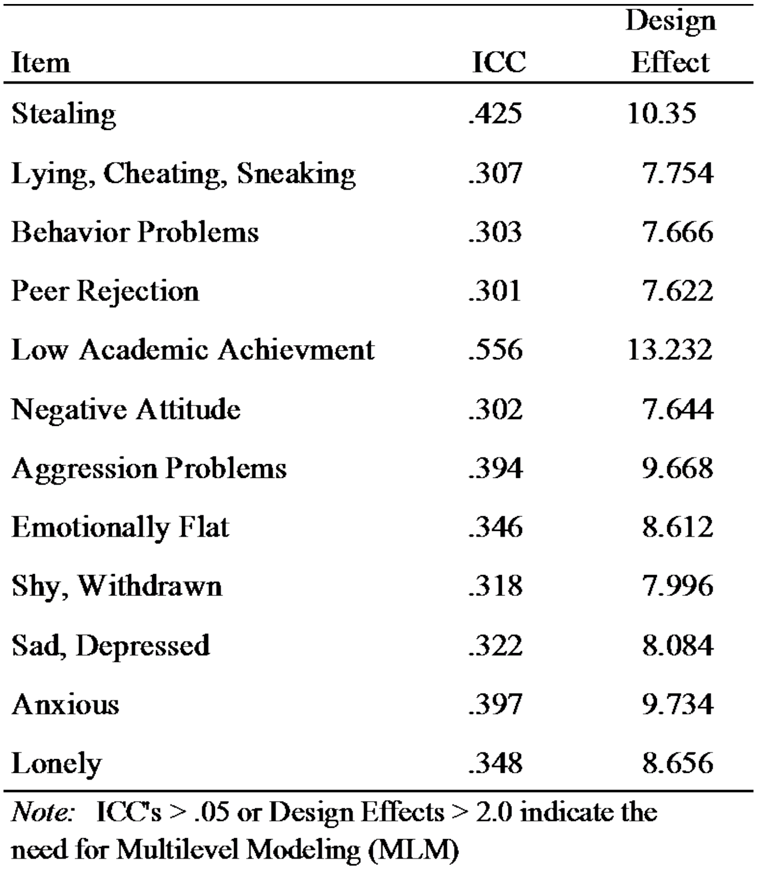
Assessing the need for multilevel modeling. A multilevel analysis was used because one assumption underlying the appropriate use of many single-level analyses is independence of errors, where similarities between observations are random (Curran, 2003). This allows for unbiased estimates of the relationship among the variables; violating this assumption can result in biased estimates of both variances and standard errors, ultimately leading to models that inaccurately represent the data. The data gathered for this study violates this assumption because it is not randomly sampled from the population. Instead, it is hierarchical, which is sometimes referred to as nested or clustered data. Hierarchical data has multiple levels, with individuals or groups nested within other groups. In schools, for example, students (level 1, or within) are nested in classes (level 2, or between), which are nested in schools (level 3, between), and so forth. For any given dependent variable measured for students (level 1), there may be some effect on that variable from them being in a particular class (level 2), or school (level 3). In cases such as this, the students' similarities on a given variable may be due primarily to effects from a higher level, rather than from chance (Rosenberg, 2009).

To determine if the clustered data needs to be analyzed using multilevel model, the variance of the dependent variables was partitioned into its within- and between-group components by calculating the Intraclass Correlation Coefficient (ICC). The ICC gives an indication of how much members of the groups (level 2) resemble one another regarding a given trait or variable (Hox, 2002). An ICC of zero indicates that there is no resemblance among members of a group for that dependent variable and a multilevel model is not necessary. An ICC greater than .05 indicates that members of the group resemble one another, potentially due to some group effect, enough variance to warrant a multilevel model. For example, an ICC of .154 shows that 15.4% of the variance is due to some aspect of the group rather than individuals.

Table 3 shows that the ICC's for the 12 items on the SRSS—IE12 range between .301 and .556 indicating that there is a multilevel effect on these dependent variables. In other words, between 30% and 55% of the variance for these dependent variables can be explained by the level 2 groups. As such, CFA models need to take into account clustering to specify an accurate model (McCoach & Black, 2012). The average cluster size of the 93 classes is 22.817, from between nine and forty-seven students.

Table 3

Intraclass Correlations and Design Effects



Item	ICC	Design Effect
Stealing	.425	10.35
Lying, Cheating, Sneaking	.307	7.754
Behavior Problems	.303	7.666
Peer Rejection	.301	7.622
Low Academic Achievment	.556	13.232
Negative Attitude	.302	7.644
Aggression Problems	.394	9.668
Emotionally Flat	.346	8.612
Shy, Withdrawn	.318	7.996
Sad, Depressed	.322	8.084
Anxious	.397	9.734
Lonely	.348	8.656

Note: ICC's > .05 or Design Effects > 2.0 indicate the need for Multilevel Modeling (MLM)

Due to the clustered nature of the data, Multilevel Confirmatory Factor Analysis (MCFA) was primarily used to analyze the data for the three research questions; all analyses were completed using Version 7.4 of Mplus. As the name implies, MCFA was used to confirm, or provide evidence of a theoretical relationship between latent (factors) and observed variables. It is appropriate for this study because of its utility (a) in the psychometric evaluation of a psychological scale in testing hypotheses concerning constructs or latent traits, (b) testing the assumptions of Cronbach's coefficient α and providing parameter estimates necessary for other reliability coefficients, and (c) evaluating factorial invariance (Brown, 2015).

In the data for this study, the 12 items on the SRSS-IE are student (level 1) variables, but the reported behaviors exhibited in a given class period may be influenced by some class-level (level 2) variables. Although the purpose of the current study was not to identify specific sources of classroom effects on student scores regarding the SRSS-IE, the MCFA approach accounts for the variance introduced from students being nested classrooms.

Choosing the appropriate parameter estimator. The default estimation of Maximum Likelihood (ML) in Mplus assumes the data is continuous and multivariate normal; however, the data for this study is ordinal, and the distribution of errors resulting from ordinal data cannot be normal due to the limited range of possible responses. Hox (2010), argues that by default, scores obtained from ordered categories, such as the SRSS-IE, fail to meet the assumption of multivariate normality. The data from the SRSS-IE, however, are traditionally highly skewed because it screens for aberrant behaviors in a general population. A Mardia's test for multivariate skew (Mardia, 1970) returned a statistic of 84.69, $p < .001$, indicating a significant difference between a symmetric normal distribution and the multivariate distribution of the scores from the SRSS-IE. Data with a large sample size, however, can result in significant differences even with small deviations from normality. Because of this, Ullman and Ullman (2006) recommend visual inspection of the data in addition to significance tests. The multivariate distribution of scores reported from the SRSS-IE is displayed in Figure 1. Thus, the assumption of multivariate normality is rejected, and if ML were used as an estimator, it would increase the probability of type I error (DiStephano, 2002).

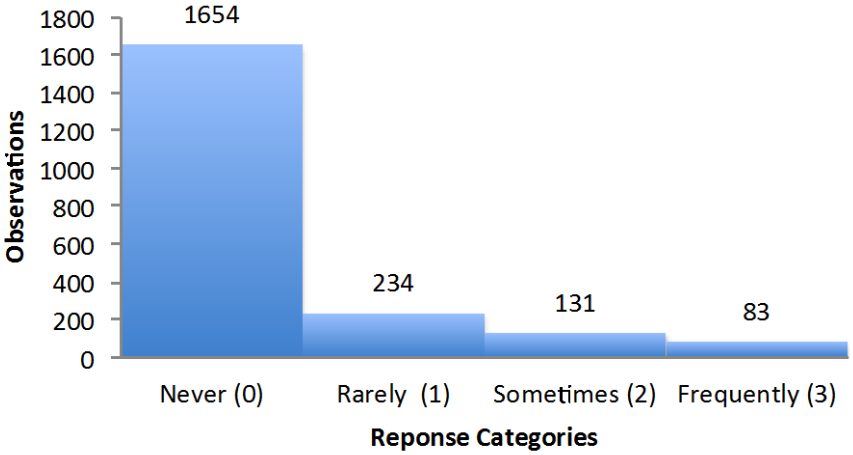


Figure 1. Multivariate distribution of SRSS-IE

An alternative to ML estimation is Robust Weighted Least Squares, also known as Weighted Least Squares Mean and Variance adjusted (WLSMV), This estimator has been shown to provide accurate estimates when *n* > 200, and is the default estimator for ordered categorical data in Mplus (Muthén, Toit, & Spisic, 1997; Rhemtulla, Brosseau-Liard, & Savalei, 2012). Therefore, all single-level models use WLSMV estimator in obtaining parameter estimates for all three questions. Because WLSMV cannot be used for testing Factorial Invariance with the student-level (level 1) grouping variable of gender, Weighted Least Squares Mean adjusted (WLSM) estimator was used to specify two-level models, and the scaling factor reported in the Mplus output was used to perform the chi-square difference test among nested models. (Satorra, 2000). Maximum Likelihood, using numeric integration for use with ordinal data, was used to estimate Factorial Invariance.

Question 1: Factor structure. Lane et al. (2013) conducted a principal component analysis (PCA) of the scores on the SRSS-IE from a middle school population. Their study found two latent traits that supported the hypothesized structure of Internalizing and Externalizing factors. The Items Stealing, Lying/Cheating/Sneaking, Behavior Problems, Low Academic Achievement, Negative Attitude, and Aggression Problems were found to load solely on the Externalizing Factor, while Emotionally Flat, Sad, Shy/Withdrawn, Anxious, and Lonely were found to load entirely on the Internalizing factor. Peer Rejection was found to load on both factors. This proposed factor structure was tested using Confirmatory Factor Analysis with data from a middle-school population using approaches that account for the clustered nature of the data.

To fit the best model, various alternates, including single-level and two-level models, general factor and two-factor models, and nested models with varying parameter constraints are considered (Brown, 2015; Kline, 2011; Peugh, 2010). Although single-leveling modeling does not provide the flexibility of estimating different parameters among the levels, Mplus can take into account clustered data in a single-level model using the *type = complex* command, where the chi-square and standard errors are adjusted to account for the clustered nature of the data (Múthen & Satorra, 1995). One of the benefits of specifying a single-level model before a two-level model its relative simplicity to fit and interpret (Dedrick & Greenbaum, 2010). Further, single-level models provide a valuable comparison to other research on the factor structure of the SRSS-IE, which employ single-level modeling. However, even with the adjusted chi-squares and standard errors, a single-level model may fail to adequately account for the clustered data, or accurately represent the factor structure.

A general two-level MCFA model can be specified by the following equation (Heck & Thomas, 2015; retaining their notation):

$$\Upsilon_{ij} = \nu_B + \Lambda_B \eta_{Bj} + \varepsilon_B + \Lambda_{Wij} \eta_{Wij} + \varepsilon_{Wij}, \quad (1)$$

where an individual i in group j's score on a given variable γ is a function of the following: γ is a vector of observed variables, υ is a vector of intercepts, Λ is a factor loading matrix, η represents random factor components, and ε is residual variance. The subscripts B and W stand for between and within, respectively. Thus, as indicated by the equation above, multilevel models decompose the variance for a specified dependent variable, or set of variables, to within-group covariance (level 1, which are students in this study) and between-group covariance (level 2—class). The within- and between-group covariance are represented separately in the model; subsequently there are more parameters to estimate. Further, the factor structure among levels may differ. While the multilevel approach is generally more complicated to model and interpret, it may reveal a factor structure that is more informative. However, with more accurate models there is the increased likelihood of fewer findings of good model fit or of statistical significance (Pedhazur & Schmelkin, 1991). Three-level models (classrooms nested in schools) will not be considered due insufficient number of school clusters.

*Model comparison, evaluation, and re-specification***.** These models were evaluated, and in certain cases re-specified based on theory and model results, with two models retained for further analysis in questions 2 and 3. Both single and multilevel models were evaluated using global and local fit. Global Fit indices indicate the extent to which the relationships among variables in the model, as shown by the model implied covariance matrix, match the relationships in the observed data, as specified by the observed covariance matrix. Global fit indices are divided into absolute and relative fit indices.

*Absolute fit indices***.** Examples of this type of fit index include the Root Mean Square Error of Approximation (RMSEA; Steiger & Lind, 1980), the Root-Mean Square Residual (RMSR) and the chi-square test of model fit. The latter is the traditional means of assessing model fit; it directly compares the difference between the model implied and observed matrices, with a significant p-value indicating poor fit. However, absolute fit indices can be biased depending on the number of observed variables, sample size (Miles & Shevlin, 2007). Further bias can come as a result of non-multivariate normal data, like that of the SRSS-IE (MacCallum, Browne, & Sugawara, 1996). While standards for good fit vary in the literature as to cut points for acceptable fit, for this study values < .10 are considered acceptable. Further, to balance the limitations of the absolute fit indices, relative fit indices were also reported to provide a more complete picture of model fit (Brown, 2015).

*Relative fit indices***.** Known as comparative or incremental fit indices, these indices compare the hypothesized model to an independent, or null model where the observed variables are uncorrelated, and there are no latent variables. Examples of these fit indices include Bentler's Comparative Fit Index (CFI; Bentler & Bonett, 1980) and the Tucker-Lewis Index, sometimes referred to as the Normed Fit Index (TLI and NFI respectively; Tucker & Lewis, 1973). Estimates of these fit indices generally range between 0.00 and 1.00; while standards for good fit vary in the literature, for this study the following cutoff values follow the guidelines set forth in Brown (2015), as follows: poor: .00-.84; acceptable: .85-.89; good: .90-.94; close:.95-1.00. While Mplus returns the fit indices listed above for the overall model, two-level fit indices for both levels are calculated for retained models using the method set forth by Ryu (2014b).

Two other fit indices worth noting are the Akaike Information Criterion (AIC; Akaike, 1974), and the Bayes Information Criterion (BIC; Schwartz, 1978). These are parsimony corrected fit indices that take into account the complexity of a model and add penalties as the number of estimated parameters increase. The AIC and BIC are useful in comparing non-nested models, with the model having the lowest value indicating better fit. Nested models were compared using the chi-square difference test.

Local fit. Examining local fit can provide more detail concerning specific parts of the model, and can inform respecifiction for a misfitting model (Kline, 2011). Local fit includes checking individual parameter estimates, such as factor loadings, intercepts, thresholds, residual variances, and standard errors to find values that are either out of range or not statistically significant. As stated earlier, based on evaluation and re-specification, two models were retained for analysis for questions 2 and 3.

Question 2: Estimating internal consistency reliability. Estimating reliability is an important aspect of providing evidence of validity, but Cronbach's coefficient α, the most commonly reported reliability coefficient, is often reported without verifying that its assumptions have been met. Violations of these assumptions can lead to α either overestimating or underestimating the reliability of a scale. Cronbach's α assumes tau-equivalence and uncorrelated errors (Komaroff, 1997; Lord & Novick, 1968; Raykov, 2010); CFA provides a framework by which these assumptions can be verified. In order to meet the assumption of tau-equivalence: (a) the factor model must be congeneric, meaning that each observed variable must load on only one latent variable (no cross-loadings), and (b) the factor loadings for a given latent trait can be constrained to be equal without significantly reducing model fit. Composite reliability (ω: McDonald, 1970, 1999) and Maximal reliability (H; Conger, 1980) were used to estimate reliability where the assumptions of Cronbach's α were violated (Geldhof, Preacher, & Zyphur, 2014). Composite reliability (ω) is estimated as

$$\omega = \frac{\left(\sum_{i=1}^{k} \lambda_i\right)^2}{\left(\sum_{i=1}^{k} \lambda_i\right)^2 + \sum_{i=1}^{k} \theta_{ii}} \tag{2}$$

where λ_i and θ_{ii} are the factor loading and unique variance of item *i,* respectively. Maximal reliability (H) was estimated as

$$H = \left(1 + \frac{1}{\sum_{i=1}^{k} \frac{\varrho_i^2}{1 - \varrho_i^2}}\right)^{-1} \tag{3}$$

where ϱ_i is the standardized factor loading for item *i*. These methods of estimating reliability were used to report reliability coefficients for the retained models. For the two-level model, parameter estimates were used to calculate reliability on both levels.

Question 3: Evaluating measurement invariance/equivalence between males and females. Question 3 addresses the issue of Measurement Invariance/ Equivalence of the SRSS—IE12 between male and female students. This comparison is important because previous research studies have shown that males tend to be nominated more often for externalizing behaviors while internalizing behaviors tend to be considered more feminine (Frank, 2000; Hoffman, 2009; Young et al., 2010). It is important to distinguish whether the difference is due to the actual trait or due to bias inherent in the screener itself.

Measurement Invariance/ Equivalence is sometimes referred to as Factorial Invariance or Differential Item Functioning (DIF). For this study, the term Factorial Invariance is used because the underlying goal is to examine whether or not the scale, or its individual items, are biased against a subpopulation. Factorial Invariance is estimated on several different levels (Widaman & Reise, 1997; see also Meredith, 1993, Meredith & Horn 2001). Configural invariance tests the overall model structure between groups, with factor loadings (except for the marker variable), factor variances, and thresholds are freely estimated. Acceptable fit statistics resulting from this model indicate that the overall model is equivalent between groups. Metric invariance, also known as weak invariance, restricts the configural model by equating the factor loadings, the first threshold, and the second threshold of the marker variable. Finally, scalar invariance, also known as strong invariance, is further restricted by equating all thresholds. At each stage, a significance test is performed to test if model fit has significantly decreased; a non-significant result indicates evidence of invariance. While other levels of invariance such as strict and partial invariance appear in the literature, this study confined analysis to configural, metric, and scalar.

Examining factorial invariance is relatively well established for single-level models; this approach is readily extended to two-level models when the grouping variable is on level 2. This means that all observations for a cluster belong to exactly one group. For this to apply to this study, there could be all boys and all girl classes, but no mixed-gender classes. However, since the level-one variable of gender is both found in any given class, extending the standard model will not work. (Asparouhov & Múthen, 2012; Ryu, 2014b, 2015). Unfortunately, approaches for exploring factorial invariance when the grouping variable is on level 1 are not as well established. As recently as 2012, Kim et al. (2012) concluded that testing for factorial invariance with a level-1 variable was not feasible. Since that time, three approaches have been put forward as possible solutions.

Ryu (2014a) has shown that there is a feasible method for addressing Factorial Invariance with level 1 groups by estimating parameters using Múthen's Maximal Likelihood estimator (MUML). However, this approach assumes the data is multivariate normal and that the between level has an *n* greater than 100. As the data for this study violates both of these assumptions, this approach is not used.

In an unpublished working paper, Asparouhov and Múthen (2012) propose using a Factor Mixture Model (FMM), generally used for Latent Class Analysis (LCA), to investigate differences in level-1 groups in a two-level model. This specific approach is also passed over. While this estimation method has been shown to work with non-normal multivariate data, their specific application to factorial invariance is too narrow.

Kim, Yoon, Wen, Luo, and Kwok (2015) broaden the Factor Mixture Model approach used by Asparouhov and Múthen to test for configural, metric, and scalar invariance with a level-one grouping variable. This approach was used to test for factorial invariance in this study. Although they further propose using Multiple Indicators Multiple Causes (MIMIC) estimation to test for invariance among the individual items of a scale, it was deemed unstable for this data set, as parameter estimates returned are well outside the normal range.

Summary

To provide evidence of validity concerning internal structure, this study employed a cross-sectional design with Multilevel Confirmatory Factor Analysis to further explore the evidence for validity in regards to the scores obtained from the Student Risk Screening Scale—Internalizing/Externalizing. This was studied by first confirming the hypothesized factor structure with both single-level and two-level models. Next reliability was estimated using the approach suggested by Geldhof et al. (2014). Finally, factorial invariance was explored using MIMIC method proposed by Kim et al. (2015). The following section reports the results of the data analysis.

CHAPTER 4: RESULTS

In the previous section, the method for conducting the study has been reported. This chapter will report the results of the data analysis for each of the following research questions:

To what extent do SRSS-IE scores from a middle-school population show evidence of:

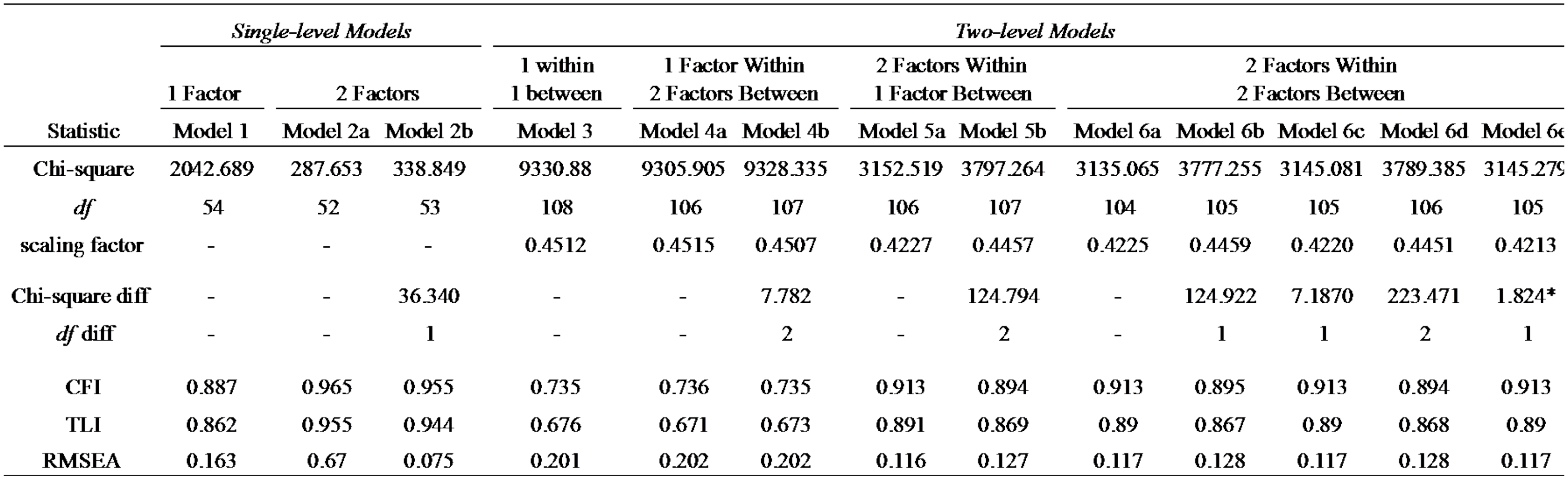
1. A two-factor structure with Peer Rejection loading on both factors, as proposed by Lane et al. (2013)?
2. Internal consistency reliability?
3. Measurement equivalence/invariance (ME/I) between males and females?

Question 1: Factor Structure

Table 4 displays global fit indices for 13 alternate models including single and two-level models, including general factor and two-factor models. The two-factor models 2, 4, and 5 also include nested models, *b* nested within *a*, with Peer Rejection loading only on *Externalizing,* and Peer Rejection loading on both factors, respectively. Models 1 through 6a are diagramed in Figure 2. Due to the number of models, not all are diagramed. Model 6 includes a two-factor model on both levels, making the comparison of nested models with Peer Rejection loadings and cross-loadings more complicated. As such, it includes models *b* through *e* nested in model *a*. Corresponding diagrams of these nested models are found in Figure 3.

Table 4

Analysis of Fit Among 13 Alternate Models



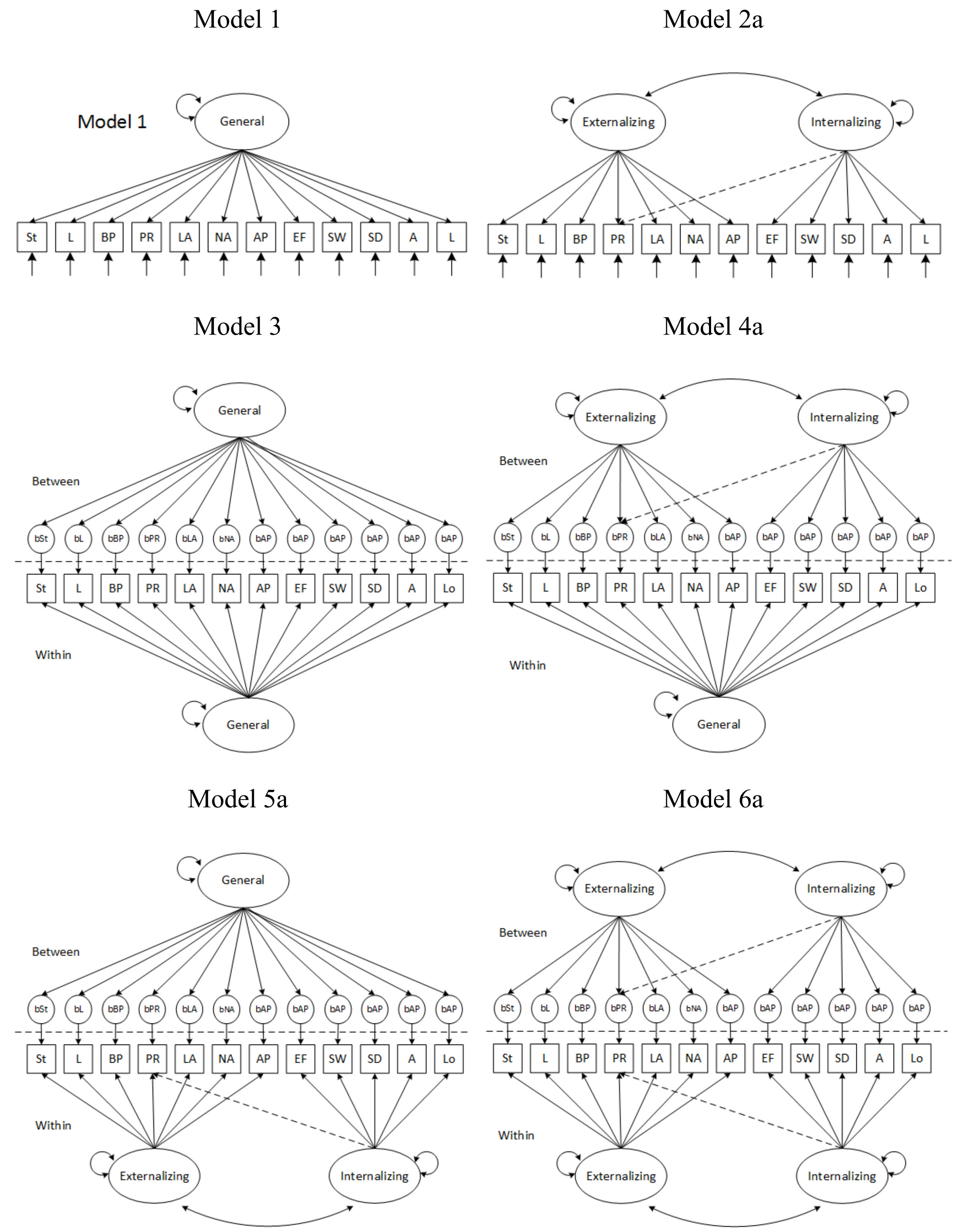
	Single-level Models			*Two-level Models*									
	1 Factor	2 Factors		1 within 1 between	1 Factor Within 2 Factors Between		2 Factors Within 1 Factor Between		2 Factors Within 2 Factors Between				
Statistic	Model 1	Model 2a	Model 2b	Model 3	Model 4a	Model 4b	Model 5a	Model 5b	Model 6a	Model 6b	Model 6c	Model 6d	Model 6e
Chi-square	2042.689	287.653	338.849	9330.88	9305.905	9328.335	3152.519	3797.264	3135.065	3777.255	3145.081	3789.385	3145.279
df	54	52	53	108	106	107	106	107	104	105	105	106	105
scaling factor	-	-	-	0.4512	0.4515	0.4507	0.4227	0.4457	0.4225	0.4459	0.4220	0.4451	0.4213
Chi-square diff	-	-	36.340	-	-	7.782	-	124.794	-	124.922	7.1870	223.471	1.824*
df diff	-	-	1	-	-	2	-	2	-	1	1	2	1
CFI	0.887	0.965	0.955	0.735	0.736	0.735	0.913	0.894	0.913	0.895	0.913	0.894	0.913
TLI	0.862	0.955	0.944	0.676	0.671	0.673	0.891	0.869	0.89	0.867	0.89	0.868	0.89
RMSEA	0.163	0.67	0.075	0.201	0.202	0.202	0.116	0.127	0.117	0.128	0.117	0.128	0.117

Note: All models (a) have Peer Rejection loading on both factors, while models (b) have Peer Rejection loading only on the Externalizing factor. The differences in model 6 are as follows: 6a Peer Rejection loads on both factors on both levels, 6b Peer Rejection loads on both factors on level 2, but only Externalizing on level 1, 6c Peer Rejection loads on both factors on level 1, but only on Externalizing on level 2, 6d Peer rejection does not load on the Internalizing factor on either level, 6e Peer rejection loads on both factors on level 1, but does not load on Externalizing on Level 2.
*indicates the nested model has significantly better fit than the comparison model, p<.001.

Single-level models. As noted earlier, these single-level models are specified with WLSMV estimator and the *type = complex* argument in Mplus to adjust the chi-square and standard errors in accordance with the clustered nature of the data. General factor and two-factor models are specified as models 1 and 2 respectively. The general factor model (model 1) shows poor fit from all global indices. Model 2 is further split into a comparison model with Peer Rejection loading on both factors (2a), and a nested model with Peer Rejection loading only on Externalizing (2b). Although both models show good fit, Model 2a fits significantly better than the nested model, with fit indices showing close fit: Δx^2 = 38.609 (1), p < .001. All parameters of the models were statistically significant.

Two-level models. Testing the factor structure and cross-loading with two-level allows for many more factor structure possibilities. These models are specified with the WLSM estimator so that nested models can be compared with the chi-square difference test using the Satorra-Bentler scaling factor. The general factor model (3) shows poor fit from all indices. Models 4 and 5 explore hypothesized models where the student and classroom-level factor structures differ. Model 4 has one factor on the within level (student), and two factors on the between level (classroom), and model 5 has this arrangement reversed. As with the single-level models, both models 4 and 5 are further split with comparison models where Peer Rejection loads on both factors (4a, 5a), and nested models where Peer Rejection loads only on Externalizing (4b, 5b). Models 4a and 4b both show poor fit overall; however, loading Peer Rejection on both factors significantly improves the fit of 4a over 4b: Δx^2 = 7.782 (1), p < .05. Model 5b borders on good fit; Model 5a, however has significantly better fit, with some fit indices breaking into the good fit range: Δx^2 = 124,794, p < .001. All parameters of the models were plausible and statistically significant.

Model 6a has two factors on both levels with Peer Rejection loading on both factors on both levels (see Figure 2). Because of the many possible variations of cross-loadings that can be tested, Figure 3 shows the alternative models nested within 6a. Tested against the nested models where Peer Rejection is not allowed to cross-load on Internalizing on the within, between, and both levels, respectively, 6a has significantly better fit: Δx^2 = 124.922, p < .001; Δx^2 = 7.187, p < .007; Δx^2 = 223.471, p < .001. Further inspection of the parameter estimates for model 6a, however, revealed that Peer Rejection's loading on Externalizing was insignificant on the classroom level: λ_{24} = .391, p = .622. Subsequently, a fourth nested model, 6e, was specified with Peer Rejection only loading on Internalizing on the Classroom level. Unlike the other nested models, the fit of 6e was not significantly worse than that of 6a: Δx^2 = 1.824 (1), p = .1767. In order to further investigate the fit of the two-level models, level-specific fit is reported in Table 5 for three models: 5a, 6a, and 6e. These three models were chosen for several reasons; first, while they have the best fit of all the two-level models analyzed for this study, the fit estimates generally fall in the mediocre to good range. Examining level-specific fit may reveal misfit on one of the two levels, and thereby inform future re-specified models. Second, all three models have almost identical overall fit estimates. While model 6a and 6e have been directly compared, model 5a cannot because it is not nested in either model. Thus, the level-specific fit may provide guidance as to which model should be retained for further analysis in this study.


Model 1

Model 2a

Model 3

Model 4a

Model 5a

Model 6a

Figure 2. Selected model diagrams.

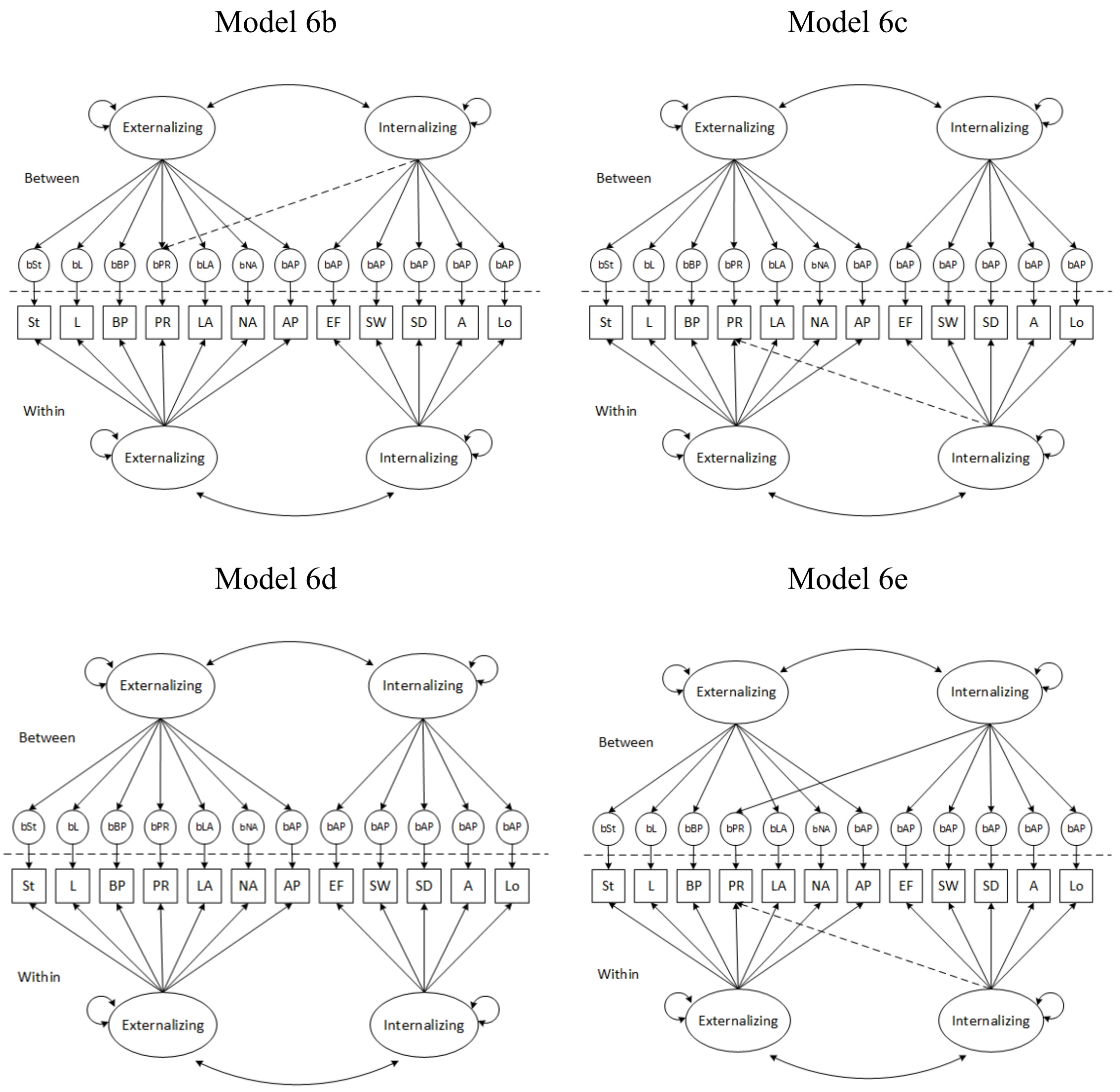
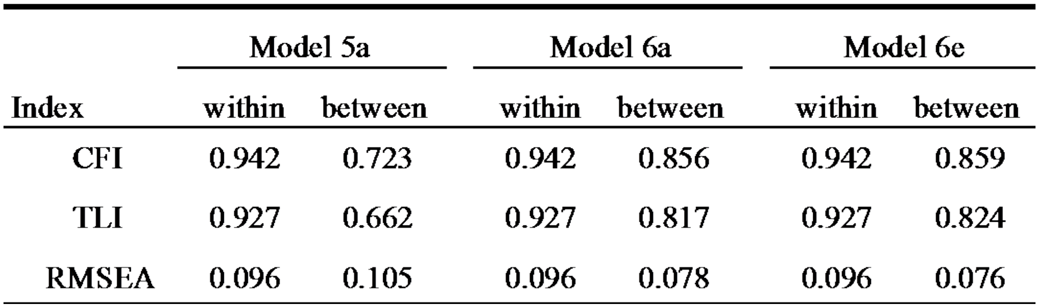


Figure 3. Alternative models nested within Model 6a.

Notice that the within-level fit indices are identical among models because they share the same within-level factor structure. Between-level results indicate poor to mediocre fit across all three models, with a potentially greater degree of misfit with model 5a. Further, within-level results indicate good fit at best (as opposed to close fit). As noted earlier, fewer findings of significance, or in this case good fit, are too be expected as statistical models increase in complexity, as is the case with multi-level models compared to single level models (Pedhazur & Schmelkin, 1991). However, the misfit of these models may also suggest that there is more research to be done to fit more accurate models. While a purely exploratory approach is beyond the scope of this paper, the information here may prove a needed starting point.

Table 5

Multi-Level Global Fit of Three Competing Models



	Model 5a		Model 6a		Model 6e	
Index	**within**	**between**	**within**	**between**	**within**	**between**
CFI	0.942	0.723	0.942	0.856	0.942	0.859
TLI	0.927	0.662	0.927	0.817	0.927	0.824
RMSEA	0.096	0.105	0.096	0.078	0.096	0.076

Based on the global fit, local fit, and level-specific fit, two models, 2a and 6e are retained for further analysis in questions 2 and 3. Model 2a and the student-level portion of 6e confirm the two-factor model proposed by Lane et al. (2013). Model 6e further appears to have better classroom-level fit than model 5a and is more parsimonious than model 6a. Table 6 displays the parameter estimates of the two retained models.

Question 2: Estimating Internal Consistency Reliability

This section estimates the internal consistency of the two subscales of the SRSS-IE. As described in the previous chapter, for Cronbach's coefficient α to serve

Table 6

Parameter Estimates for Retained Models

	Model 2a: Single-level Two-factor with Crossoading						Model 6e: Two-level Two-factor with Crossloading on Student Level								
							Student Level (within)			Classroom Level (between)					
	Loadings			Thresholds			Loadings			Loadings			Thresholds		
item	Ext	Int	SE	0,1	1,2	2,3	Ext	Int	SE	Ext	Int	SE	0,1	1,2	2,3
Stealing	1.000	-	0.000	1.437	2.110	-	1.000	-	0.000	1.000	-	0.000	3.433	5.033	-
Lying, Cheating, Sneaking	1.076	-	0.040	0.707	1.205	1.745	1.184	-	0.091	1.088	-	0.206	1.680	2.964	4.309
Behavior Problems	0.997	-	0.048	0.728	1.135	1.669	1.025	-	0.974	0.813	-	0.187	1.548	2.470	3.655
Peer Rejection	0.513	0.509	0.065/.087	0.897	1.481	2.110	0.391	0.527	).036/.056	-	0.566	0.156	1.375	2.348	3.355
Low Academic Achievment	0.928	-	0.046	0.253	0.710	1.172	0.666	-	0.057	0.949	-	0.235	0.179*	1.130	2.100
Negative Attitude	1.039	-	0.048	0.708	1.208	1.693	0.775	-	0.057	0.766	-	0.163	1.217	2.163	3.058
Aggression Problems	1.010	-	0.044	1.053	1.605	2.088	0.814	-	0.059	0.963	-	0.220	2.118	3.230	4.141
Emotionally Flat	-	1.000	0.000	0.875	1.361	1.885	-	1.000	0.000	-	1.000	0.000	1.463	2.307	3.154
Shy, Withdrawn	-	0.865	0.039	0.460	0.910	1.382	-	0.767	0.051	-	0.721	0.114	0.575	1.276	2.009
Sad, Depressed	-	1.149	0.069	0.935	1.474	2.048	-	1.831	0.168	-	1.477	0.237	2.244	3.641	5.028
Anxious	-	0.927	0.033	0.998	1.503	2.039	-	0.813	0.058	-	0.829	0.142	1.525	2.413	3.289
Lonely	-	1.094	0.067	0.971	1.485	2.068	-	1.672	0.131	-	1.357	0.156	2.168	3.416	4.757
Variance	0.650	0.628	0.628	-	-	-	2.651	1.071	.378/.116	1.454	0.886	.664/.281	-	-	-
Correlation	0.523		0.037	-	-	-	0.336		0.021	0.834		0.049	-	-	-

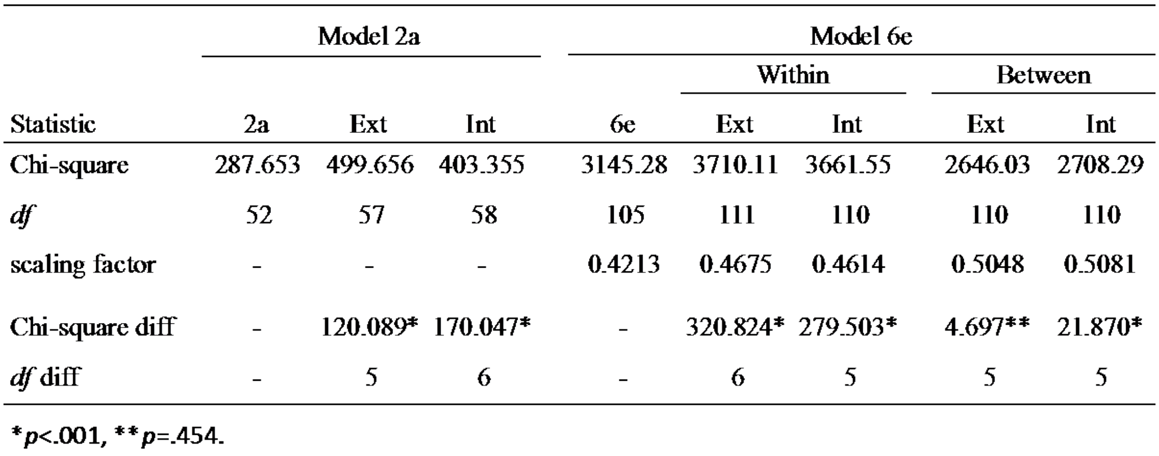
Note: **Threshold parameters are not estimated for the within-level of multilevel models.**

***indicates a non-significant parameter estimate**

as an unbiased estimator of reliability, scores from the SRSS-IE must meet the assumptions of essential tau-equivalence and there must be no correlated errors for a given subscale (Raykov, 1997, 2010). Neither of the retained models (2a and 6e) specifies any correlated errors, but both contain one cross loading, which violated the assumption of tau-equivalence. To test for the other stipulation of tau-equivalence, each factor variance for each subscale in either model is set to 1, and the factor loadings are constrained to be equal. This nested model is then tested against the original model using a chi-square difference test. A significant result indicates that the nested model has significantly worse fit, thus suggesting that the assumption of equal factor loadings is also violated.

Table 7

Analysis of Fit with Constrained Factor Loadings



	Model 2a			Model 6e				
					Within		Between	
Statistic	2a	Ext	Int	6e	Ext	Int	Ext	Int
Chi-square	287.653	499.656	403.355	3145.28	3710.11	3661.55	2646.03	2708.29
df	52	57	58	105	111	110	110	110
scaling factor	-	-	-	0.4213	0.4675	0.4614	0.5048	0.5081
Chi-square diff	-	120.089*	170.047*	-	320.824*	279.503*	4.697**	21.870*
df diff	-	5	6	-	6	5	5	5

**p*<.001, ***p*=.454.

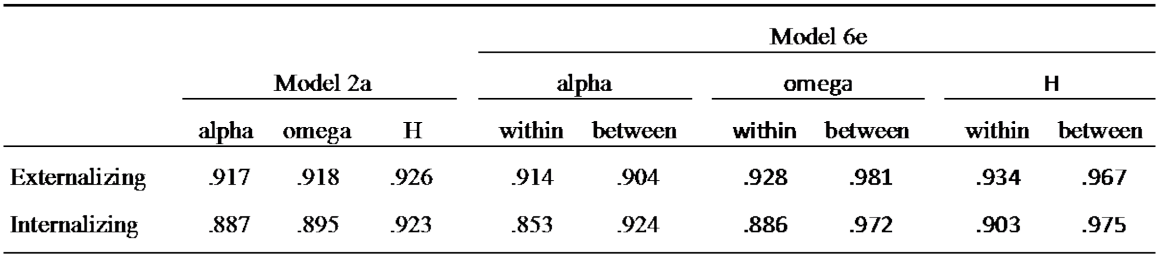
Table 7 shows model fit of the original (comparison) models compared to nested models (where the factor loadings are constrained to be equal) to test for tau-equivalence. With the exception of the Externalizing factor on the classroom level of model 6e, all other loading-equated models show significantly worse fit than the comparison model.

Thus, neither model 2a nor 6e meet the assumptions of tau-equivalence, meaning that Cronbach's coefficient α is likely to provide biased estimates of reliability.

Reliability of the scores obtained through the SRSS-IE is estimated using composite reliability ω and maximal reliability H, for each subscale, as described earlier. Further, level-specific estimates are included for model 6e. Table 8 displays the result of the reliability estimates, including estimate from α and two-level α, for comparison.

Table 8

Internal Consistency Reliability Estimates



				Model 6e					
	Model 2a			alpha		omega		H	
	alpha	omega	H	within	between	within	between	within	between
Externalizing	.917	.918	.926	.914	.904	.928	.981	.934	.967
Internalizing	.887	.895	.923	.853	.924	.886	.972	.903	.975

Notice that the estimates from ω and H are comparable, while Cronbach's coefficient α appear to slightly underestimate reliability. Generally speaking, ω and H estimate higher reliability with the Externalizing subscale; further, reliability on the classroom-level (between) is generally higher than the student level (within). Despite the slight variation in reliability estimates, both models indicate a high level of internal reliability for each of the subscales (Nunnally & Bernstein, 1994).

Question 3: Evaluating Measurement Equivalence/Invariance

This section compares the factor structure of models 2a and 6e between females (n = 1042) and males (n =1080). Table 9 displays the fit indices and significance tests of model 2a for configural, metric, and scalar invariance.

Table 9

Model 2a Single-Level Factorial Invariance

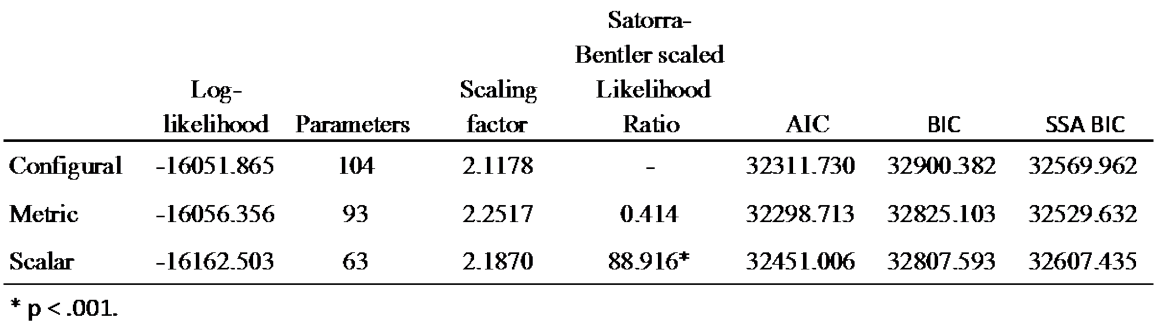
	Chi-Square value	*df*	Chi Square Difftest Value	*df* difference	*p*-value	CFI	TLI	RMSEA
Configural	464.367	104	-	-	-	.953	0.940	.057
Metric	469.827	115	14.876	11	0.188	.953	0.946	.054
Scalar	452.004	136	15.935	21	0.773	.958	0.960	.047

The indices for the configural model show close fit, indicating that the same factor structure holds across groups. Because the standard approach for testing differences in the chi-square values of nested models does not work when using WLSMV or WLSM estimators, the *Difftest* command was used to test for significant model misfit between Configural and Metric, and Metric and Scalar models. The *p*-values for both significance tests were above .05, indicating that the model fit was not significantly different from one model to another. Thus, the structure of model 2a shows evidence of strong invariance between females and males.

When examining factorial invariance using the Factor Mixture Model for 6e, convergence was not met using the default settings in Mplus. To reach convergence, the Robust Maximum Likelihood (MLR) estimator was used, the convergence criterion was relaxed from .001 to 0.1, and the number of integration points were reduced. Because the MLR estimator was used, a scaling factor was used to conduct the Likelihood Ratio Test (LRT; Satorra & Bentler, 2000). Results of Factorial Invariance testing are reported in Table 10.

Table 10

Model 6e Two-Level Factorial Invariance



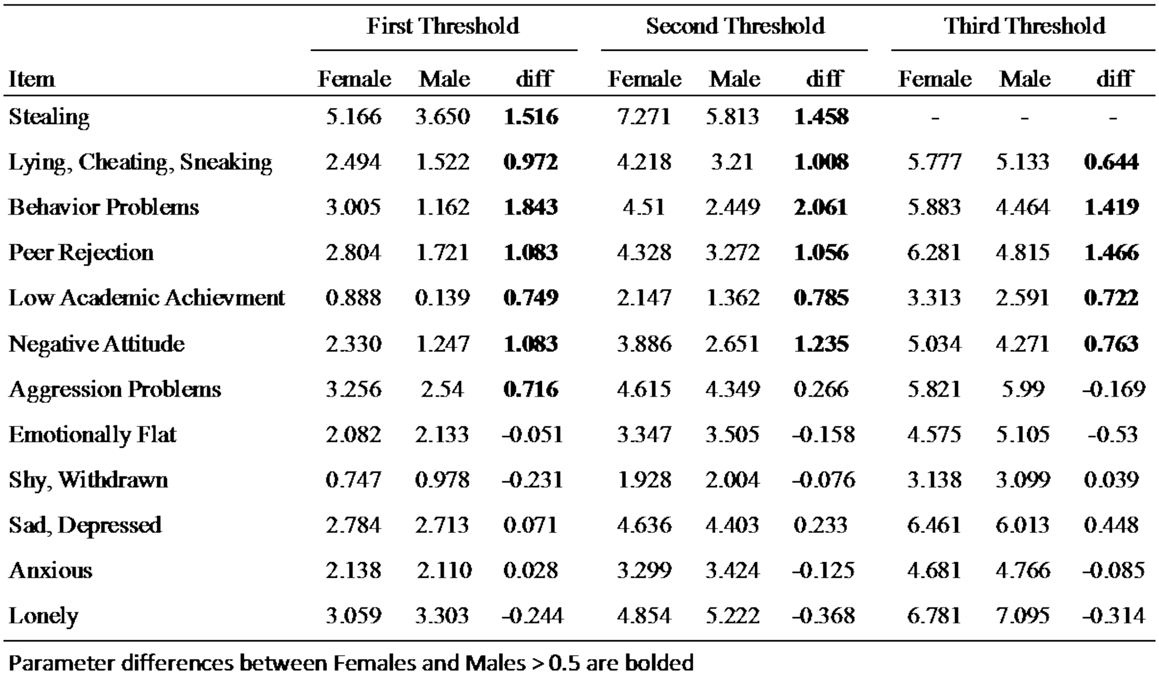
	Log-likelihood	Parameters	Scaling factor	Satorra-Bentler scaled Likelihood Ratio	AIC	BIC	SSA BIC
Configural	-16051.865	104	2.1178	-	32311.730	32900.382	32569.962
Metric	-16056.356	93	2.2517	0.414	32298.713	32825.103	32529.632
Scalar	-16162.503	63	2.1870	88.916*	32451.006	32807.593	32607.435

* p < .001.

The two-level model suggests that the scores from the SRSS-IE exhibit metric, or weak, invariance, but not scalar, or strong invariance. This means that item threshold estimates between females and males are significantly different. The threshold indicates the point where a rater (in this case the teacher) has an equal probability of selecting adjacent response categories. For *k* response categories there are *k* - 1 thresholds. The scale for the SRSS-IE, for example, has four categories labeled 0 to 3, and therefore has three thresholds: the first is between categories 0 and 1, the second between categories 1 and 2, and the third between categories 2 and 3. As stated above, the threshold represents the point where a rater has an equal probability of choosing either of the two adjacent categories. To further investigate the specific sources of invariance, the number and magnitude of nonequivalent threshold parameters are explored. Table 11 shows the parameter estimates for females and males, as well as the difference between them. Differences greater than 0.5 are bolded.

Table 11

Threshold Parameter Estimates and Differences Between Females and Males



	First Threshold			Second Threshold			Third Threshold		
Item	**Female**	**Male**	**diff**	**Female**	**Male**	**diff**	**Female**	**Male**	**diff**
Stealing	5.166	3.650	**1.516**	7.271	5.813	**1.458**	-	-	-
Lying, Cheating, Sneaking	2.494	1.522	**0.972**	4.218	3.21	**1.008**	5.777	5.133	**0.644**
Behavior Problems	3.005	1.162	**1.843**	4.51	2.449	**2.061**	5.883	4.464	**1.419**
Peer Rejection	2.804	1.721	**1.083**	4.328	3.272	**1.056**	6.281	4.815	**1.466**
Low Academic Achievment	0.888	0.139	**0.749**	2.147	1.362	**0.785**	3.313	2.591	**0.722**
Negative Attitude	2.330	1.247	**1.083**	3.886	2.651	**1.235**	5.034	4.271	**0.763**
Aggression Problems	3.256	2.54	**0.716**	4.615	4.349	0.266	5.821	5.99	-0.169
Emotionally Flat	2.082	2.133	-0.051	3.347	3.505	-0.158	4.575	5.105	-0.53
Shy, Withdrawn	0.747	0.978	-0.231	1.928	2.004	-0.076	3.138	3.099	0.039
Sad, Depressed	2.784	2.713	0.071	4.636	4.403	0.233	6.461	6.013	0.448
Anxious	2.138	2.110	0.028	3.299	3.424	-0.125	4.681	4.766	-0.085
Lonely	3.059	3.303	-0.244	4.854	5.222	-0.368	6.781	7.095	-0.314

Parameter differences between Females and Males > 0.5 are bolded

The results indicate that significant non-equivalence between gender groups appears to occur with items that load almost entirely on the Externalizing factor, with males more likely to be rated in a higher category on the scale than females for an observed behavior. The negative values indicate instances where females are more likely to be rated for a higher category for a given observed behavior, and that these instances occur almost entirely with items that load on the Internalizing factor.

Summary

Data analysis has been conducted in an attempt to answer three questions concerning evidence for the validity of the internal structure of the SRSS-IE. The three questions deal with factor structure, reliability, and factorial invariance; the results have been reported. Further discussion of the results will take place in the following section.

CHAPTER 5: DISCUSSION

This chapter summarizes the findings of the research, discusses possible interpretations and implication of the analysis, describes limitations of the study, and provides suggestions for future research.

The purpose of this study was to examine evidence of validity concerning the internal structure of the SRSS-IE in a middle school population. This is relevant because although there is ample evidence of its validity with elementary-aged populations, the evidence for validity in a secondary setting is not as well established; as such, it may not be correctly identifying students who are most in need of timely interventions for Emotional and Behavioral Disorders (EBD).

To address this issue, three questions concerning the factor structure, reliability, and factorial invariance of the scores from the SRSS-IE were analyzed and reported. Results indicate that the hypothesized two-factor model with Peer Rejection loading on both factors has good fit in a single-level model. A two-level model has been proposed with acceptable fit, with the same student-level structure as the single-level model, and a two-factor classroom-level model with Peer Rejection loading on Internalizing rather than Externalizing. Internal consistency estimates indicate high internal consistency reliability for all subscales. Using composite reliability ω and maximal reliability H, estimates range from .886 to .981. Finally, both retained models suggest strong factorial invariance between the scores of females and males.

Discussion

While the results from the single-level analysis are relatively straightforward, the two-level model analysis returned mixed results that complicate the interpretation.

Single-level models. If modeled on a single level, data obtained from the SRSS-IE and subsequent analysis make a strong case for validity based on its internal structure. The evidence suggests that the two-factor solution with Peer Rejection loading on both factors proposed by Lane et al. (2013) exhibits good fit with data drawn from a middle school population. The measure shows high reliability and produces scores that are invariant between females and males. According to this model, teacher and administrators have at least preliminary evidence that the SRSS-IE measures what it purports to measure. This model is also attractive because of its ease of interpretation, as compared to the two-level model.

There are, however, caveats to relying solely on this model. First, the approach of the three questions proposed by this study is primarily confirmatory; the single-level models test an already defined structure and its psychometric properties. This defined structure is based on the theory that there are two hypothetical constructs labeled *Externalizing* and *Internalizing* as proposed by Drummond (1994), and Lane (2012, 2013). However, there may be other alternate factor structures based on other theory that equally or better account for relationships among the variables (Kline, 2011). For example, as noted in the review of the literature, there are at least two competing factor structures for the Strengths and Difficulties Questionnaire (SDQ). Some researchers found a three-factor solution (Di Riso et al., 2010; Matsuishi et al., 2008; Van Roy et al., 2008; Yao et al., 2009), while other research showed a five-factor solution to be best fitting (R. Goodman, 2001; Smedje et al., 1999; Thabet et al., 2000). It is possible that the 12 items from the SRSS-IE could be structured differently as well, but only the prevailing theory of two factors has been tested. Further, even within the bounds of the current theory of two factors (with Peer Rejection loading on both), there may be other cross loadings or correlated errors that make theoretical sense and contribute to better model fit.

Second, single-level modeling may misrepresent the data. The Intraclass Correlations (ICC) for these data range between .301 and .556. Recall that the ICC indicates the extent to which the data violate the assumption of independence of observations, and an ICC with a value greater than .05 is a general indicator that this assumption has been violated. While the single-level models in this study adjust the chi-square and standard errors using the *type = complex* command, it is unlikely that with such high ICC's that this approach adequately accounts for clustered nature of the data.

Ignoring clustering of the data has been an issue with research concerning the internal structure evidence for validity among universal screeners in general. None of the studies that examine the factor structures of any of the screeners in the review of the literature mention testing the assumption of independence of observations, report the ICC, or employ multilevel modeling as part of their analysis. Granted, in some instances, the data for these studies was not clustered or did not have sufficient statistical power. However, the data from the majority of the reviewed research was drawn from contexts where the individuals were nested in classes or other groups, and many of them had sufficient power to conduct multilevel modeling.

Two-level models. The two-level models in this study account for the clustered data by decomposing the variance into within-group and between-group parts. While the general approach to modeling may be appropriate for the data, the results themselves are not as straight-forward as those from the single-level analysis.

*Multilevel fit***.** First, there is an issue with fit. As mentioned above, the approach for this study focused primarily on testing a structure already defined in previous research at the expense of exploring other possible structures. While it is not possible to directly compare single-level and two-level models with significance tests, the overall fit statistics for the best-fitting two-level models never approached the close fit range. Whereas the relative fit indices for the single level models range from .862 to .995 (from mediocre to close fit), the overall relative fit for the two-level models range from .67 to .91 (from poor to good).

The level-specific fit indices shed further light on the source of the misfit. The student-level (within) portion of the model has decent, if not ideal, fit. But the classroom-level (between) fit results are mediocre *at best*. The factor structure of the classroom-level portion of the two-level model has not been explored in the literature previous to this study. While model 6e was retained because it had the best fit among the nested models, further research is needed to establish a better classroom-level (between) model, and possibly even better student-level (within) models.

Specifying this part of the model is important because further exploration and confirmation of classroom-level (between) portion of the model could yield valuable information to schools in dealing with EBD. Although the SRSS-IE is meant to provide student-level information, with interventions directed at individuals, the classroom-level model may provide direction regarding appropriate interventions and teacher aids needed on a classroom basis. Further, while the student-level model exhibits good fit, a better fitting model would be more informative as to what these latent emotional and behavioral constructs look like in middle school, and provide valuable information as to what next steps would be most effective when dealing with at-risk students.

While the overall model fit may be improved by allowing different parameters to be freely estimated, model misfit may be due to the items themselves. Take, for example, the item *Stealing.* Out of the 2,122 students, only nine individuals were reported to have frequently been observed engaging in this behavior, eight of which were male. But such a low count with a middle school population may be the result of developmental and environmental differences from elementary students. On the developmental side, *Stealing* may be a better indicator for elementary students because impulsivity tends to decrease with age (Steinberg et al., 2008). Further, there are vast differences in the expectations and environment between middle and elementary schools which may also influence results in universal screening (Lane & Carter, 2006). For example, middle school students see multiple teachers for 45 to 60 minutes per day, as opposed to elementary students who stay with a single teacher for the majority of a school day. Due to such developmental and environmental factors, items such as *Stealing* may not provide the same level of information with older children. This is a theoretical issue that is not directly addressed in this study, but the misfit of the two-level model at least opens the door for further discussion and research on the matter.

The results of both single-level and multi-level models indicate that Peer Rejection loads on both factors on the within-level (student level). Theoretical reasons for this cross-loading have not been discussed in previous literature, and there are several possible reasons it hasn't loaded on a single construct. One possible explanation is that the meaning of this item is ambiguous, leading some teachers to interpret it as the rated student being rejecting their peers, and leading others to interpret it as peers rejecting the student. It may also be that teachers are interpreting it solely as the latter of these two options; if so this item is potentially problematic as it is the only item on the entire scale that is not rating a student's personal behavior. Instead, it is rating other student's behavior toward the rated student. While more research needs to be done to more fully explore how teachers are interpreting this item, this study indicates that this item either needs to be clarified, or dropped from the scale altogether.

*Two-level reliability***.** Model fit also has implications regarding estimating reliability and evaluating invariance, as both procedures require acceptable model fit to return unbiased results (Zyphur, Kaplan, & Christian, 2008). In regards to internal consistency, reliability estimates are higher than that found in previous research. Using H and ω, estimates among the subscales and levels ranged between .89 and .98. Previous research in middle schools estimated reliability using Cronbach's coefficient α of the subscales between .76-.89 (Ennis et al., 2011; Kalberg, Lane, & Menizes, 2007; Oakes et al., 2010). Data that violates tau-equivalence often over-estimates reliability. For this data, α seems to have underestimated reliability, which is usually an indication of correlated errors.

A closer look at the estimates between subscales and levels provides further insight into the reliability of the scores. One interesting finding is that even though the classroom-level (between) portion of the two-level model has poor to mediocre fit, its reliability estimates were often higher than those of the better fitting student-level (within) portion. This suggests that model fit and scale reliability describe different properties of the scores from a given scale; but that is not to say that one does not in some ways affect the other. Both composite reliability (ω) and maximal reliability (H) rely on factor loadings to compute internal consistency coefficients. Different models produce different parameter estimates that can result in varying reliability estimates. For example, the classroom-level portion of model 6e does not mirror the student-level portion; with the former, Peer Rejection loads only on Internalizing. Had Peer Rejection loaded on both factors, as found in the student-level portion, there would be another parameter to include in the reliability estimate, perhaps altering it. Thus, while the classroom-level reliability estimates are higher than the student-level estimates, they should be taken as preliminary estimates. Future research may fit a better model with different parameters that could change the reliability estimates.

*Multilevel factor invariance***.** An interesting finding of this study is the failure of the two-level model to meet the standards for scalar or strong invariance. As mentioned earlier, this is a deviation from the single-level model, which is found to be invariant across item thresholds. A visual inspection of the two-level parameter estimates for females and males reveals that the non-equivalence lies almost entirely with items that load on the Externalizing factor. The threshold parameters for females are higher than for males. This means that teachers have less of a propensity to rate female students with a higher response category on a given item than a male. Conversely, the lower threshold for males indicates more of a propensity to rate a male student with a higher response category on a given item. Overall, it means that teachers are more likely to give a higher rating on an externalizing behavior to a male than a female.

Teacher nomination of males more frequently for externalizing behaviors is consistent with existing research (Frank, 2000; Hoffman, 2009; Young et al., 2010).

However, interpreting this finding is potentially problematic. It is unclear from this analysis why teachers are more likely to rate males with a higher response category than females. It may indicate that the instrument is better at addressing the externalizing construct for males, and less efficient at doing so for females. Another interpretation is that males naturally exhibit externalizing behaviors more frequently than females.

This gender difference is not the case with items that load on the Internalizing factor. An examination of the threshold estimate differences between females and males for Internalizing items are less than 0.5; further, difference values are both positive and negative. In sum, results from the two-level model suggest that teachers tend to rate males differently than females on Externalizing items, but do not systematically discriminate between males and females regarding Internalizing items. This is an important property of the scale for teachers and administrators to understand when interpreting the scores from the SRSS-IE. It suggests that either males are potentially being identified as being at-risk when they are not, or that females are not being identified who are at risk, or both.

While these results suggest that the Externalizing subscale is not invariant, these results should be taken as preliminary and not definitive for several reasons. First, methods for dealing with factorial invariance using a grouping variable from level 1 have not been fully established in the literature, much less in standard practice. Thus there may be problems with the method of estimation rather than with the instrument itself. Second, methods testing factorial invariance with multilevel models have been primarily fleshed out using data that is multivariate normal; because of the severe skew of the SRSS-IE data, results might be biased and unreliable. This issue manifests itself during the data analysis where on several occasions, the model failed to converge or exhibited other estimation problems such as unreasonable parameter estimates. With time and more research, better estimation methods for factorial invariance for data that violate multivariate normality may emerge. Third, the grouping variable of gender is measured on level 1 (student level/within), but in MCFA, the item thresholds are estimated on level 2 (classroom level/between). Because the between-level portion of the model specified in this study has only mediocre fit, the between-level comparison of thresholds may be mediocre as well. In other words, because a close-fitting classroom-level portion of the model has not been estimated yet, the factorial invariance results may not be trustworthy and would be different with a better fitting model.

Limitations

There are several limitations to this study. First, one drawback of cross-sectional designs, in general, is the limited scope. Because the data were only collected on one occasion, there is no evidence to support whether or not the internal structure of the SRSS-IE is invariant over time. It may be that teacher perceptions of students, and therefore their ratings of student behavior change over time. It may also be that student behavior changes over time.

Second, unlike elementary schools where students spend their entire school day with a single teacher, middle school students’ time is divided equally among six or seven 45 to 60-minute classes. This is a potential source of variance in the data: a student’s behavior in Math may be very different than in Physical Education, and the teachers’ ratings will reflect only what they observe in their limited 45 minute-a-day window. This study only requested that the first-period teachers rate their students for that period. It may be that the scores would vary had a different teacher from a different class rated that particular student.

Third, the population from which the data was drawn is predominantly from an urban, Caucasian sample. As such, results may not generalize to other races or ethnicities, or to those in rural settings.

Finally, the data itself is not multivariate-normal, and the results may be biased as a result, as indicated in the previous section. While analytic procedures were used that attempt to take into account the non-normality of the data, some analyses, such as two-level factorial invariance, have not been fully developed, thus producing suspect results.

Future Directions

Several of the limitations in the previous section concerning this study can be addressed through further research. First, a longitudinal study, including test-retest reliability, could address the issue of time invariance. Second, different designs for collecting data from more than one teacher for a single student, or accounting for the inter-rater variance may provide a more accurate picture of student behavior. Such designs may include G-theory or the Many-facets Rasch model. Third, the results of this study should be confirmed with diverse populations to examine the models in varying contexts.

There are several other suggestions for future research that stem from the previous discussion of the results of this study. Primary among them is the need to fully explore the two-level factor structure and estimate a close-fitting model, in particular on the classroom-level (between). As multilevel modeling continues to grow as a standard for analyzing hierarchical data. Further development of unbiased estimates of factorial invariance for level-1 variables when the scale items are ordinal is also needed. The estimation of such a model will provide better estimates of reliability and evaluation of factorial invariance. Item Response Theory (IRT) is another approach for evaluating internal structure; while the factor analysis with categorical data in Mplus is essentially a 2-parameter logistic model, there may be other models in the family of IRT that fit the data better. Further, IRT offers other categories for Measurement Equivalence/Invariance or Differential Item Functioning (DIF) that may yield different and potentially more accurate results (Millsap & Yun-Tein, 2004).

Conclusion

Ultimately, this study sought to answer this question: is the SRSS-IE an appropriate means to screen for EBD in middle schools? The answer to this question may be different for each school depending on a host of varying factors, and certainly cannot be definitively resolved by a single study. However, inasmuch as this study helps to inform teachers, administrators, parents, and other stakeholders as to the strengths and weakness of the SRSS-IE, it has fulfilled its purpose. The results of this study are promising, in terms of providing evidence for the validity of the SRSS-IE, and it adds to a growing body of research for this instrument. But more than the validation of a specific instrument, in its small way this study furthers the work of identifying students at-risk for EBD, in the hopes of quicker and more accurate identification. The hope is that as more schools implement effective screening policies and intervention strategies, the lives of many more young people will be vastly improved.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *Transactions on Automatic Control*, *19*(6), 716–723.

Allen-DeBoer, R. A., Malmgren, K. W., & Glass, M. E. (2006). Reading instruction for youth with emotional and behavioral disorders in a Juvenile correction facility. *Behavioral Disorders*, *32*(1), 18–28.

American Educational Research Association (2014). *Standards for educational and psychological testing.* Washington D.C.: AERA Publications.

Asparouhov, T., & Múthen, B. (2012). Multiple group multilevel analysis. *Mplus Webnotes.* Retreived from http://www.statmodel.com/examples/webnotes/webnote16.pdf

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*(3), 588–606.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York, NY: The Guiford Press.

Bullis, M., & Yovanoff, P. (2006). Idle hands: Community employment experiences of formerly incarcerated youth. *Journal of Emotional and Behavioral Disorders*, *14*(2), 71–85.

Caldarella, P., Young, E. L., Richardson, M. J., Young, B. J., & Young, K. R. (2008). Validation of the systematic screening for behavior disorders in middle and junior high school. *Journal of Emotional and Behavioral Disorders*, *16*(2), 105–117. doi:10.1177/1063426607313121

Chin, J. K., Dowdy, E., & Quirk, M. P. (2012). Universal screening in middle school:

Examining the Behavioral and Emotional Screening System. *Journal of Psychoeducational Assessment*, *31*(1), 53–60. doi:10.1177/0734282912448137

Conger, A. J. (1980). Maximally reliable composites for unidimensional measures. *Educational and Psychological Measurement*, *40*(2), 367–375.

Conners, C. K. (1997). Conners' Parent Rating Scale--Revised (L). North Tonawanda, NY: Multi-Health Systems.

Curran, P. J. (2003). Have multilevel models been structural equation models all along?. Multivariate Behavioral Research, 38(4), 529-569.

Dedrick, R. F., & Greenbaum, P. E. (2010). Multilevel confirmatory factor analysis of a scale measuring interagency collaboration of children's mental health agencies. *Journal of Emotional and Behavioral Disorders*, *20*(10), 1–14. doi:10.1177/1063426610365879

Dever, B. V., Mays, K. L., Kamphaus, R. W., & Dowdy, E. (2012). The factor structure of the BASC-2 Behavioral and Emotional Screening System teacher form, child/adolescent. *Journal of Psychoeducational Assessment*, *30*(5), 488–495. doi:10.1177/0734282912438869

Di Riso, D., Salcuni, S., Chessa, D., Raudino, A., Lis, A., & Altoè, G. (2010). The Strengths and Difficulties Questionnaire (SDQ). Early evidence of its reliability and validity in a community sample of Italian children. *Personality and Individual Differences*, *49*(6), 570–575. doi:10.1016/j.paid.2010.05.005

Dickey, W. C., & Blumberg, S. J. (2004). Revisiting the factor structure of the Strengths and Difficulties Questionnaire: United States, 2001. *Journal of the American Academy of Child and Adolescent Psychiatry*, *43*(9), 1159–67.

doi:10.1097/01.chi.0000132808.36708.a9

DiStephano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling*, *9*(3), 327–346.

Dowdy, E., Chin, J. K., & Quirk, M. P. (2013). Preschool screening: An examination of the Behavioral and Emotional Screening System Preschool teacher form (BESS Preschool). *Journal of Psychoeducational Assessment*, *31*(6), 578–584. doi:10.1177/0734282913475779

Dowdy, E., Chin, J. K., Twyford, J. M., & Dever, B. V. (2011). A factor analytic investigation of the BASC-2 Behavioral and Emotional Screening System Parent Form: Psychometric properties, practical implications, and future directions. *Journal of School Psychology*, *49*(3), 265–80. doi:10.1016/j.jsp.2011.03.005

Dowdy, E., Dever, B. V, Distefano, C., & Chin, J. K. (2011). Screening for emotional and behavioral risk among students with limited English proficiency. *School Psychology Quarterly*, *26*(1), 14–26. doi:10.1037/a0022072

Dowdy, E., Twyford, J. M., Chin, J. K., Distefano, C. A., Kamphaus, R. W., & Mays, K. L. (2011). Factor structure of the BASC – 2 Behavioral and Emotional Screening System student form. *Psychological Assessment*, *23*(2), 379–387. doi:10.1037/a0021843

Drummond, T. (1994). *The Student Risk Screening Scale (SRSS)*. Grants Pass, OR: Josephine County Mental Health Program.

Dunn, D. M., & Dunn, L. M. (2007). *Peabody Picture Vocabulary Test* (4th ed.). Circle Pines, MN: Pearson Assessments.

Durlak, J. A., & Weissberg, R. P. (2007). *The impact of after-school programs that* *promote personal and social skills*. Chicago, ILL: Collaborative for Academic, Social, and Emotional Learning.

Egger, H. L., & Angold, A. (2006). Common emotional and behavioral disorders in preschool children: Presentation, nosology, and epidemiology. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *47*(3-4), 313–337. doi:10.1111/j.1469-7610.2006.01618.x

Elias, M. J., & Arnold, H. (Eds.). (2006). *The educator's guide to emotional intelligence and academic achievement: Social and emotional learning in the classroom* (1st ed.). Thousand Oaks, CA: Corwin Press.

Ennis, R. P., Lane, K. L., & Oakes, W. P. (2011). Score reliability and validity of the Student Risk Screening Scale: A psychometrically sound, feasible tool for use in urban elementary schools. *Journal of Emotional and Behavioral Disorders*, *20*(4), 241–259. doi:10.1177/1063426611400082

Ezpeleta, L., Granero, R., De la Osa, N., Penelo, E., & Domènech, J. M. (2013). Psychometric properties of the Strengths and Difficulties Questionnaire in 3-year-old preschoolers. *Comprehensive Psychiatry*, *54*(3), 282–91. doi:10.1016/j.comppsych.2012.07.009

Frank, E. (Ed.). (2000). *Gender and its effects on psychopathology* (1st ed.). Washington D.C.: American Psychiatric Press, Inc.

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, *19*(1), 72–91. doi:10.1037/a0032138

Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology*, *45*(2), 117–135. doi:10.1016/j.jsp.2006.05.005

Goodman, A., & Goodman, R. (2009). Strengths and Difficulties Questionnaire as a dimensional measure of child mental health. *Journal of the American Academy of Child and Adolescent Psychiatry*, *48*(4), 400–3. doi:10.1097/CHI.0b013e3181985068

Goodman, A., & Goodman, R. (2011). Population mean scores predict child mental disorder rates: Validating SDQ prevalence estimators in Britain. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *52*(1), 100–8. doi:10.1111/j.1469-7610.2010.02278.x

Goodman, A., Lamping, D. L., & Ploubidis, G. B. (2010). When to use broader internalising and externalising subscales instead of the hypothesised five subscales on the Strengths and Difficulties Questionnaire (SDQ): Data from British parents, teachers and children. *Journal of Abnormal Child Psychology*, *38*(8), 1179–91. doi:10.1007/s10802-010-9434-x

Goodman, R. (1999). The extended version of the Strengths and Difficulties Questionnaire as a guide to child psychiatric caseness and consequent burden. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *40*(5), 791–9. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10433412

Goodman, R. (2001). Psychometric properties of the Strengths and Difficulties Questionnaire. *Journal of the American Academy of Child and Adolescent Psychiatry*, *40*(11), 1337–45. doi:10.1097/00004583-200111000-00015

Gresham, F. M., Lane, K. L., & Lambros, K. M. (2000). Comorbidity of conduct problems and ADHD: Identification of "fledgling psychopaths." *Journal of Emotional and Behavioral Disorders*, *8*(2), 83–93. doi:10.1177/106342660000800204

Hagquist, C. (2007). The psychometric properties of the self-reported SDQ – An analysis of Swedish data based on the Rasch model. *Personality and Individual Differences*, *43*(5), 1289–1301. doi:10.1016/j.paid.2007.03.022

Haynes, A., Gilmore, L., Shochet, I., Campbell, M., & Roberts, C. (2013). Factor analysis of the self-report version of the Strengths and Difficulties Questionnaire in a sample of children with intellectual disability. *Research in Developmental Disabilities*, *34*(2), 847–854. doi:10.1016/j.ridd.2012.11.008

Heck, R. H., & Thomas, S. L. (2015). *An introduction to multilevel modeling techniques* (3rd ed.). New York, NY: Routledge.

Hill, C. R., & Hughes, J. N. (2007). An examination of the convergent and discriminant validity of the Strengths and Difficulties Questionnaire. *School Psychology Quarterly : The Official Journal of the Division of School Psychology, American Psychological Association*, *22*(3), 380–406. doi:10.1037/1045-3830.22.3.380

Hoffman, D. M. (2009). Reflecting on social emotional learning: A critical perspective on trends in the United States. *Review of Educational Research*, *79*(2), 533–556. doi:10.3102/0034654308325184

Hox, J. (2002). *Multilevel analysis: Techniques and applications* (1st ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Hox, J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). Florence, KY: Routledge.

Jee, S. H., Halterman, J. S., Szilagyi, M., Conn, A., Alpert-Gillis, L., & Szilagyi, P. G. (2011). Use of a brief standardized screening instrument in a primary care setting to enhance detection of social-emotional problems among youth in foster care. *Academic Pediatrics*, *11*(5), 409–13. doi:10.1016/j.acap.2011.03.001

Kalberg, J. R., Lane, K. L., Driscoll, S., & Wehby, J. (2010). Systematic screening for emotional and behavioral disorders at the high school level: A formidable and necessary task. *Remedial and Special Education*, *32*(6), 506–520. doi:10.1177/0741932510362508

Kamphaus, R. W., Distefano, C., Dowdy, E., Eklund, K., & Dunn, A. R. (2010). Determining the presence of a problem : Comparing two approaches for detecting youth behavioral risk. *School Psychology Review*, *39*(3), 395–407.

Kamphaus, R. W., & Reynolds, C. R. (2007). *BASC-2 Behavioral and Emotional Screening System manual*. Circle Pines, MN: Pearson.

Kauffman, J. W. (2001). *Characteristics of emotional and behavioral disorders of children and youth* (7th ed.). Colombus, OH: Merrill.

Kim, E. S., Kwok, O., & Yoon, M. (2012). Testing factorial invariance in multilevel data: A Monte Carlo study. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(2), 250–267.

Kim, E. S., Yoon, M., Wen, Y., Luo, W., & Kwok, O. (2015). Within-level group factorial invariance with multilevel data: Multilevel factor mixture and multilevel MIMIC models. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(4), 603–616. doi:10.1080/10705511.2014.938217

King, K., Reschly, A. L., & Appleton, J. J. (2012). An examination of the validity of the

Behavioral and Emotional Screening System in a rural elementary school: Validity of the BESS. *Journal of Psychoeducational Assessment*, *30*(6), 527–538. doi:10.1177/0734282912440673

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: The Guilford Press.

Komaroff, E. (1997). Effect of simultaneous violations of essential tau-equivalence and uncorrelated error on coefficient alpha. *Applied Psychological Measurement*, *21*(4), 337–348. doi:10.1177/01466216970214004

Kovacs, M., & Devlin, B. (1998). Internalizing disorders in childhood. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *39*(1), 47–63. doi:10.1111/1469-7610.00303

Landrum, T. J., Tankersley, M., & Kaufman, J. M. (2003). What is special about special education for students with emotional or behavioral disorders? *The Journal of Special Education*, *37*(3), 148–156.

Lane, K. L., Bruhn, A. L., Eisner, S. L., & Robertson Kalberg, J. (2010). Score reliability and validity of the Student Risk Screening Scale: A psychometrically sound, feasible tool for use in urban middle schools. *Journal of Emotional and Behavioral Disorders*, *18*(4), 211–224. doi:10.1177/1063426609349733

Lane, K. L., & Carter, E. W. (2006). Supporting transition-age youth with and at risk for emotional and behavioral disorders at the secondary level: A need for further inquiry. *Journal of Emotional and Behavioral Disorders*, *14*(2), 66–70. doi:10.1177/10634266060140020301

Lane, K. L., Little, M. A., Casey, A. M., Lambert, W., Wehby, J., Weisenbach, J. L., &

Phillips, A. (2008). A comparison of systematic screening tools for emotional and behavioral disorders. *Journal of Emotional and Behavioral Disorders*, *17*(2), 93–105. doi:10.1177/1063426608326203

Lane, K. L., Menzies, H. M., Oakes, W. P., Lambert, W., Cox, M., & Hankins, K. (2012). A validation of the Student Risk Screening Scale for Internalizing and Externalizing Behaviors: Patterns in rural and urban elementary schools. *Behavioral Disorders*, *37*(4), 244–270.

Lane, K. L., Oakes, W. P., Carter, E. W., Lambert, W. E., & Jenkins, A. B. (2013). Initial Evidence for the Reliability and Validity of the Student Risk Screening Scale for Internalizing and Externalizing Behaviors at the Elementary Level. *Assessment for Effective Intervention*, *37*(2), 99–122. doi:10.1177/1534508413489336

Lane, K. L., Oakes, W. P., Harris, P. J., Menzies, H. M., Cox, M., & Lambert, W. (2012). Initial evidence for the reliability and validity of the Student Risk Screening Scale for Internalizing and Externalizing Behaviors at the elementary level. *Behavioral Disorders*, *37*(2), 99–122.

Lane, K. L., Parks, R. J., Kalberg, J. R., & Carter, E. W. (2007). Systematic screening at the middle school level: Score reliability and validity of the Student Risk Screening Scale. *Journal of Emotional and Behavioral Disorders*, *15*(4), 209–222.

Lane, K. L., Kalberg, J. R., Lambert, E. W., Crnobori, M., & Bruhn, A. L. (2009). A comparison of systematic screening tools for emotional and behavioral disorders: A replication. *Journal of Emotional and Behavioral Disorders*, *18*(2), 100–112. doi:10.1177/1063426609341069

Lane, K. L., Kalberg, J. R., Parks, R. J., & Carter, E. W. (2008). Student Risk Screening

Scale: Initial evidence for score reliability and validity at the high school level. *Journal of Emotional and Behavioral Disorders*, *16*(3), 178–190. doi:10.1177/1063426608314218

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*(1), 85–106. doi: http://doi.org/10.1037//1076-898X.8.2.99

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling of fit involving a particular measure of model. *Psychological Methods*, *13*(2), 130–149. doi:10.1037/1082-989X.1.2.130

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with application. *Biometrika*, *57*(3), 519–530.

Matsuishi, T., Nagano, M., Araki, Y., Tanaka, Y., Iwasaki, M., Yamashita, Y., Kakuma, T. (2008). Scale properties of the Japanese version of the Strengths and Difficulties Questionnaire (SDQ): A study of infant and school children in community samples. *Brain & Development*, *30*(6), 410–5. doi:10.1016/j.braindev.2007.12.003

McCoach, D. B., & Black, A. C. (2012). Introduction to estimation issues in multilevel modeling. New Directions for Institutional Research, 2012(154), 23-39. doi:10.1002/ir.20012

McCrory, C., & Layte, R. (2012). Testing competing models of the Strengths and Difficulties Questionnaire's (SDQ's) factor structure for the parent-informant instrument. *Personality and Individual Differences*, *52*(8), 882–887.

doi:10.1016/j.paid.2012.02.011

McDonald, R. P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. British Journal of Mathematical and Statistical Psychology, 23(1), 1-21.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahway, NJ: Erlbaum.

Menzies, H. M., & Lane, K. L. (2011). Validity of the Student Risk Screening Scale: Evidence of predictive validity in a diverse, suburban elementary setting. *Journal of Emotional and Behavioral Disorders*, *20*(2), 82–91. doi:10.1177/1063426610389613

Miles, J., & Shevlin, M. (2007). A time and place for incremental fit indices. *Structural Equation Modeling*, *42*(5), 869–874.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*(3), 479–515. doi:10.1207/S15327906MBR3903

Morris, R. J., Shah, K., & Morris, Y. P. (2002). Internalizing behavior disorders. In Lane K. L., Gresham F. M., O'Shaughnessy T. E. (Eds.), *Interventions for children with or at risk for emotional and behavior disorders* (pp. 82–91). Allyn and Bacon.

Muthén, B. O., du Toit, S. H. C., & Spisic, D. (1997). Robust inference using Weighted Least Squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. *Unpublished Manuscript*. doi:10.2139/ssrn.201668

Múthen, B., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, *25*(1995), 267–316. doi:10.2307/271070

Nunnally, I. H., & Bernstein, J. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Oakes, W. P., Wilder, K. S., Lane, K. L., Powers, L., Yokoyama, L. T. K., O'Hare, M. E., & Jenkins, A. B. (2010). Psychometric properties of the Student Risk Screening Scale: An effective tool for use in diverse urban elementary schools. *Assessment for Effective Intervention*, *35*(4), 231–239. doi:10.1177/1534508410379796

Ollendick, T. H., & King, N. J. (1994). Fears and their level of interference in adolescents. *Behaviour Research and Therapy*, *32*(6), 635–8.

Pedhazur, E., & Schmelkin, L. (1991). *Measurement, design, and analysis: An integrated approach*. HIlssdale, NJ: Lawrence Erlbaum Associates.

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology*, *48*(1), 85–112. doi:10.1016/j.jsp.2009.09.002

Raykov, T. (1997). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research*, *32*(4), 375–401. doi:10.1207/s15327906mbr3204

Raykov, T. (2010). Multivariate behavioral scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research*, *32*(4), 329–353.

Renshaw, T. L., Eklund, K., Dowdy, E., Jimerson, S. R., Hart, S. R., Earhart Jr, J., & Jones, C. N. (2009). Examining the relationship between scores on the behavioral and emotional screening system and student academic, behavioral, and engagement outcomes: An investigation of concurrent validity in elementary school. The California School Psychologist, 14(1), 81-88.

Reynolds, C. R., & Kamphaus, R. W. (2004). *Behavior Assessment System for Children--Second Edition (BASC-2)*. Bloomington, MN: Pearson.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373. doi:10.1037/a0029315

Richardson, M. J., Caldarella, P., Young, B. J., Young, E. L., & Young, K. R. (2009). Further validation of the Systematic Screener for Behavior Disorders in middle and junior high school. *Psychology in the Schools*, *46*(7), 605–15. doi:10.1002/pits

Rones, M., & Hoagwood, K. (2000). School-based mental health services: A research review. *Clinical Child and Family Psychology Review*, *3*(4), 223–41. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11225738

Rønning, J.A, Handegaard, B. H., Sourander, A., & Mørch, W.-T. (2004). The Strengths and Difficulties Self-Report Questionnaire as a screening instrument in Norwegian community samples. *European Child & Adolescent Psychiatry*, *13*(2), 73–82. doi:10.1007/s00787-004-0356-4

Rosenberg, S. L. (2009). Multilevel validity: Assessing the validity of school -level inferences from student achievement test data (Doctoral dissertation). Available from ProQuest Dissertations & Theses Full Text. (Accession Order No. 3352691). Retrieved from http://search.proquest.com.erl.lib.byu.edu/docview/304958022?accountid=4488

Ruchkin, V., Jones, S., Vermeiren, R., & Schwab-Stone, M. (2008). The Strengths and Difficulties Questionnaire: The self-report version in American urban and suburban youth. *Psychological Assessment*, *20*(2), 175–82. doi:10.1037/1040-3590.20.2.175

Ruchkin, V., Koposov, R., Vermeiren, R., & Schwab-Stone, M. (2012). The Strengths and Difficulties Questionnaire: Russian validation of the teacher version and comparison of teacher and student reports. *Journal of Adolescence*, *35*(1), 87–96. doi:10.1016/j.adolescence.2011.06.003

Ryu, E. (2014a). Factorial invariance in multilevel confirmatory factor analysis. British Journal of Mathematical and Statistical Psychology, 67(1), 172-194. doi:10.1111/bmsp.12014

Ryu, E. (2014b). Model fit evaluation in multilevel structural equation models. *Frontiers in Psychology*, *5*, 1–9. doi:10.3389/fpsyg.2014.00081

Ryu, E. (2015). Multiple group analysis in multilevel structural equation model across level 1 groups. *Multivariate Behavioral Research*, *50*(3), 300–15. doi:10.1080/00273171.2014.1003769

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis in multivariate statistical snalysis. In Heijmans R. D. H, Pollock D. S. G., & Satorra, A. (Eds.), *Innovations in Multivariate Statistical Analysis* (pp. 233–247). London: Kluwer Academic Publishers.

Satorra, A., & Bentler, P. (2000). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*(4), 507–514. doi:10.1007/bf02296192

Schwartz, G. E. (1978). Estimating the dimensions of a model. *Annals of Statistics*, *6*(2), 461–464.

Smedje, H., Broman, J. E., Hetta, J., & von Knorring, A. L. (1999). Psychometric properties of a Swedish version of the “Strengths and Difficulties Questionnaire.”

European Child & Adolescent Psychiatry, *8*(2), 63–70. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10435454

Squires, J., Bricker, D., & Twombly, E. (2003). *Ages and Stages Questionnaires: Social-Emotional*. Baltimore, MD: Brookes.

Steiger, J. H., & Lind, J. (1980). Statistically-based tests for the number of common factors. Paper presented at the Annual Spring Meeting of the Psychometric Society. Iowa City.

Steinberg, L., Albert, D., Cauffman, E., Banich, M., Graham, S., & Woolard, J. (2008). Developmental Psychology. *Developmental Psychology*, *44*(6), 1764–1778. doi:10.1037/a0012955

Stouthamer-Loeber, M., & Loeber, R. (2002). Lost opportunities for intervention: Undetected markers for the development of serious juvenile delinquency. *Criminal Behaviour and Mental Health : CBMH*, *12*(1), 69–82. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12357258

Thabet, A. A., Stretch, D., & Vostanis, P. (2000). Child mental health problems in Arab children: Application of the Strength and Difficulties Questionnaire. *International Journal of Social Psychology*, *46*(4), 266–280.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1–10.

Ullman, J. B., & Ullman, J. B. (2006). Structural equation modeling : Reviewing the basics and moving forward. *Journal of Personality Assessment*, *87*(1), 35–50. doi:10.1207/s15327752jpa8701

Van Leeuwen, K., Meerschaert, T., Bosmans, G., De Medts, L., & Braet, C. (2006). The

Strengths and Difficulties Questionnaire in a community sample of young children in Flanders. *European Journal of Psychological Assessment*, *22*(3), 189–197. doi:10.1027/1015-5759.22.3.189

Van Roy, B., Veenstra, M., & Clench-Aas, J. (2008). Construct validity of the five-factor Strengths and Difficulties Questionnaire (SDQ) in pre-, early, and late adolescence. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *49*(12), 1304–12. doi:10.1111/j.1469-7610.2008.01942.x

Walker, H. M., & Severson, H. (1992). *Systematic screening for behavior disorders: Techincal manual*. Longmont, Co: Sopris West.

Yao, S., Zhang, C., Zhu, X., Jing, X., McWhinnie, C. M., & Abela, J. R. Z. (2009). Measuring adolescent psychopathology: Psychometric properties of the self-report Strengths and Difficulties Questionnaire in a sample of Chinese adolescents. *The Journal of Adolescent Health : Official Publication of the Society for Adolescent Medicine*, *45*(1), 55–62. doi:10.1016/j.jadohealth.2008.11.006

Young, E. L., Sabbah, H. Y., Young, B. J., Reiser, M. L., & Richardson, M. J. (2010). Gender differences and similarities in a screening process for emotional and behavioral risks in secondary schools. *Journal of Emotional and Behavioral Disorders*, *18*(4), 225–235. doi:10.1177/1063426609338858

Zins, J. E., Weissberg, R. P., Wang, M. C., & H. J., W. (Eds.). (2004). *Building academic success on social and emotional learning: What does the research say?* New York: Teachers College Press.

Zyphur, M. J., Kaplan, S. A., & Christian, M. S. (2008). Assumptions of cross-level measurement and structural invariance in the analysis of multilevel data: Problems and solutions. *Group Dynamics: Theory, Research, and Practice*, *12*(2), 127–140. doi:10.1037/1089-2699.12.2.127

APPENDIX A:

Mplus Syntax for Model 2a

```
Title: Model 2a

Data: file = "SRSSIE12.csv";
Variable: Name= Class Grade Gender Steal Lying Behave Reject Academic
        Attitude Aggres Flat Shy Sad Anxious Lonely;
        usevar = Steal Lying Behave Reject Academic
        Attitude Aggres Flat Shy Sad Anxious Lonely;
        missing = all(-99);
        categorical = all;
        cluster = Class;            ! Class variable is used as a cluster variable

define:
if steal gt 2 then steal = 2;      ! Final two categories combined for item Steal
analysis: type=complex;            ! Adjusts chi-square and SE for clustered data
estimator = WLSMV;                 ! Default for Categorical data

model:
Ext by Steal Lying Behave Academic Reject Attitude Aggres;
Int by Flat Shy Sad Anxious Lonely Reject;

! This syntax can be modified to create models 1 and 2b also

output: sampstat stdyx; modindices(20);
```

APPENDIX B:

Mplus Syntax for Model 6e

```
Title:  Model 6e

data: file = "SRSSIE12.csv";
variable: Name= Class Grade Gender Steal Lying Behave Reject Academic
        Attitude Aggres Flat Shy Sad Anxious Lonely;
        usevar = Steal Lying Behave Reject Academic
        Attitude Aggres Flat Shy Sad Anxious Lonely;
        missing = all(-99);
        categorical = all;
        cluster = Class;
define:
if steal gt 2 then steal = 2;
analysis: type=twolevel;
Estimator=WLSM; !WLSM allows for two-level model comparison using the Satorra-
Bentler correction
model:
%Within%
wExt by Steal Lying Behave Academic Reject Attitude Aggres;
wInt by Flat Shy Sad Anxious Lonely Reject;

%Between%
bExt by Steal Lying Behave Academic Attitude Aggres;
bInt by Flat Shy Sad Anxious Lonely Reject;
! Reject loads only on Internalizing Factor for the Between model

output: sampstat stdyx;
```

APPENDIX C:

Mplus Syntax for Factorial Invariance of Model 2a

```
title: Complex 2 factor Configural

data: file = SRSSIE12.csv;
variable: Name= Class Grade Gender Steal Lying Behave Reject Academic
        Attitude Aggres Flat Shy Sad Anxious Lonely;
        usevar = Steal Lying Behave Reject Academic
        Attitude Aggres Flat Shy Sad Anxious Lonely;
        missing = all(-99);
        categorical = all;
        cluster = Class;
        Grouping = Gender (0=Male 1=Female);

define:
if steal gt 2 then steal = 2;
analysis: type=complex;
parameterization=Delta;

!Reference group (Male)  Configural model
model:
! Reference group factor loadings (Configural: free; Metric equated; Scalar equated)
Ext by   Steal@1 (L1  Lying* (L2)  Behave* (L3) Reject* (L4) Academic* (L5)
Attitude* (L6) Aggres* (L7);
Int by Flat@1 (L8)  Shy* (L9) Sad* (L10) Anxious* (L11) Lonely* (L12)
Reject* (L13);

! Reference group scale factor (Configural, Metric, Scalar: fixed to 1)
{Steal@1} (S1)  {Lying@1} (S2) {Behave@1} (S3) {Reject@1} (S4)
{Academic@1} (S5) {Attitude@1} (S6) {Aggres@1} (S7) {Flat@1} (S8) {Shy@1} (S9)
{Sad@1} (S10) {Anxious@1} (S11) {Lonely@1} (S12);

!Reference group Item Intercepts (Configural: free; Metric: first threshold of all
items equated, second threshold of marker variables equated, all other free; Scalar:
All thresholds equated)

[Steal$1*] (I1) [Lying$1*] (I2) [Behave$1*] (I3) [Reject$1*] (I4) [Academic$1*] (I5)
[Attitude$1*](I6) [Aggres$1*](I7) [Flat$1*] (I8) [Shy$1*](I9) [Sad$1*](I10)
[Anxious$1*](I11) [Lonely$1*](I12);
```

```
[Steal$2*] (I21) [Lying$2*] (I22) [Behave$2*] (I23) [Reject$2*] (I24)
[Academic$2*] (I25) [Attitude$2*] (I26) [Aggres$2*] (I27) [Flat$2*] (I28)
[Shy$2*] (I29) [Sad$2*] (I210) [Anxious$2*] (I211) [Lonely$2*] (I212);

 ![Steal$3*] (I31) does not exist because final two categories are combined
 [Lying$3*] (I32) [Behave$3*] (I33) [Reject$3*] (I34) [Academic$3*] (I35)
 [Attitude$3*] (I36) [Aggres$3*] (I37) [Flat$3*] (I38) [Shy$3*] (I39)
 [Sad$3*] (I310) [Anxious$3*] (I311) [Lonely$3*] (I312)

!Refernce group Factor means (Configural, Metric, and Scalar: set to 0)
[Ext@0] [Int@0];
!Reference group Factor Variances (Configural, Metric, and Scalar: free)
Ext*; Int*;

!Comparison Group (Female)
Model Female:
!Comparison Group Factor Loadings (Configural: free; Metric: equated, Scalar:
equated)
Ext by Steal@1 Lying* Behave* Reject* Academic* Attitude* Aggres*;
Int by Flat@1 Shy*  Sad*  Anxious* Lonely* Reject*;

!Comparison group Scale Factor (Configural: fixed to 1; Metric: free; Scalar: free)
{Steal@1} {Lying@1} {Behave@1} {Reject@1} {Academic@1} {Attitude@1}
{Aggres@1} {Flat@1} {Shy@1} {Sad@1} {Anxious@1} {Lonely@1};

!Item intercepts (Configural: free; Metric: first threshold of all items equated, second
threshold of marker variables equated, all other free; Scalar: All thresholds equated)
[Steal$1*] [Lying$1*] [Behave$1*] [Reject$1*] [Academic$1*] [Attitude$1*]
[Aggres$1*] [Flat$1*] [Shy$1*] [Sad$1*] [Anxious$1*] [Lonely$1*];

[Steal$2*] [Lying$2*] [Behave$2*] [Reject$2*] [Academic$2*] [Attitude$2*]
[Aggres$2*] [Flat$2*] [Shy$2*] [Sad$2*] [Anxious$2*] [Lonely$2*];

![Steal$3*] (I31) does not exist because final two categories are combined
[Lying$3*] [Behave$3*] [Reject$3*] [Academic$3*] [Attitude$3*] [Aggres$3*]
[Flat$3*] [Shy$3*] [Sad$3*] [Anxious$3*][Lonely$3*];

!Comparison group Factor means (Configural: fixed to 0; Metric: free, and Scalar:
free)
[Ext@0] [Int@0];
!Comparison group Factor Variances (Configural, Metric, and Scalar: free)
Ext*; Int*;
output: sampstat   stdyx  modindices(4);
```

APPENDIX D:

Mplus Syntax for Factorial Invariance of Model 6e

```
Title: 6e Factorial Invariance

data: file = "SRSSIE12.csv";
variable: Name= Class Grade Gender Steal Lying Behave Reject Academic
        Attitude Aggres Flat Shy Sad Anxious Lonely;
        usevar = Steal Lying Behave Reject Academic
        Attitude Aggres Flat Shy Sad Anxious Lonely;
        categorical = all;
        cluster = Class;
        class = c(2);                  !class variable has two groups
        knownclass = c(gender=0 1);   !converts to latent to observed

define:
if steal gt 2 then steal = 2;
analysis: type=twolevel mixture;  !calls or mixture modeling
estimator = mlr;                  !employs Robust Maximum Likelihood
integration=montecarlo(500);      !reduces the number of integration points
mconv = .1;                       !relaxes the convergence criteria

model:
%Within%!
! Within reference group Factor loadings (Configural: free; Metric, Scalar: equated)
%overall%
wExt by Steal Lying Behave Academic Reject Attitude Aggres;
wInt by Flat Shy Sad Anxious Lonely Reject;

%c#2%
! Within comparison group Factor loadings (Configural: free; Metric, Scalar:
equated)
wExt by Lying* Behave* Academic* Reject* Attitude* Aggres*;
wInt by Shy* Sad* Anxious* Lonely* Reject*;

%Between%
%overall%
bExt by Steal Lying Behave Academic Attitude Aggres;
bInt by Flat Shy Sad Anxious Lonely Reject;

%c#1%
```

```
! Between Reference group Factor Loadings (Configural, Metric, Scalar: equated)
! Between reference group Factor Means (Configural, Metric, Scalar: Set to 0)
! Between reference group Factor Variance (Configural, Metric, Scalar: equated with
comparison group)
[bExt@0]; !bExt* (V1);
[bInt@0]; !bInt* (V2);

! Between reference group Item Intercepts of Marker variables (Configural, Metric,
! Scalar: equated)
[Steal$1*] (I1) [Steal$2*] (I21) [Flat$1*] (I8) [Flat$2*] (I28) [Flat$3*] (I38);

%c#2%
! Between comparison group Factor Loadings (Configural, Metric, Scalar: equated)
! Between comparison group Factor Means (Configural, Metric, Scalar: Free)
! Between reference group, Factor Variance (Configural, Metric, Scalar: equated with
comparison group)
!bExt (V1);
!bInt (V2);

! Between comparison group Item Intercepts of Marker variables (Configural,
! Metric, Scalar: equated) All other variable intercepts (Configural, Metric: Free;
! Scalar: equated)
[Steal$1] (I1) [Lying$1*] [Behave$1*] [Reject$1*] [Academic$1*]
[Attitude$1*] [Aggres$1*] [Flat$1] (I8) [Shy$1*] [Sad$1*] [Anxious$1*] [Lonely$1*];

[Steal$2] (I21) [Lying$2*] [Behave$2*] [Reject$2*] [Academic$2*] [Attitude$2*]
[Aggres$2*] [Flat$2] (I28) [Shy$2*] [Sad$2*] [Anxious$2*] [Lonely$2*];

![Steal$3*] (I31);
[Lying$3*] [Behave$3*] [Reject$3*] [Academic$3*] [Attitude$3*] [Aggres$3*]
[Flat$3] (I38) [Shy$3*] [Sad$3*] [Anxious$3*] [Lonely$3*];

output: stdyx;
```