2016-12-01

An Examination of the Psychometric Properties of the Student Risk Screening Scale for Internalizing and Externalizing Behaviors: An Item Response Theory Approach

Sara E. Moulton
Brigham Young University

An Examination of the Psychometric Properties of the Student Risk Screening Scale for Internalizing and Externalizing Behaviors: An Item Response Theory Approach

Sara E. Moulton

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

Doctor of Philosophy

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

Educational Inquiry, Measurement, and Evaluation

Brigham Young University

ABSTRACT

An Examination of the Psychometric Properties of the Student Risk Screening Scale for Internalizing and Externalizing Behaviors: An Item Response Theory Approach

Sara E. Moulton
Educational Inquiry, Measurement, and Evaluation, BYU
Doctor of Philosophy

This research study examined the psychometric properties of the *Student Risk Screening Scale for Internalizing and Externalizing Behaviors* (SRSS-IE) using Item Response Theory (IRT) methods among a sample of 2,122 middle school students. The SRSS-IE is a recently revised screening instrument aimed at identifying students who are potentially at risk for emotional and behavioral disorders (EBD). There are two studies included in this research. Study 1 utilized the Nominal Response and Generalized Partial Credit models of IRT to evaluate items from the SRSS-IE in terms of the degree to which the response options for each item functioned as intended by the scale developers and how well those response options discriminated among students who exhibited varying levels of EBD risk. Results from this first study indicated that the four response option configurations of the items on the SRSS-IE may not adequately discriminate among the frequency of externalizing and internalizing behaviors demonstrated by middle school students. Recommendations for item response option revisions or scale scoring revisions are discussed in this study.

In study 2, differential item functioning (DIF) and differential step functioning (DSF) methods were used to examine differences in item and response option functioning according to student gender variables. Additionally, test information functions (TIFs) were used to determine whether preliminary recommendations for cut scores differ by gender. Results of this second study indicate that two of the items on the SRSS-IE systematically favor males over females and one item systematically favors females over males. Additionally, examination of TIFs demonstrated different degrees of measurement precision at various levels of theta for males and females on both the externalizing and internalizing constructs. Implications of these results are discussed in relation to possible revisions of the SRSS-IE items, cut scores, or scale scoring procedures.

Keywords: Student Risk Screening Scale, emotional and behavioral disorders, universal screening, Item Response Theory, Nominal Response Model, differential item functioning, cut scores

ACKNOWLEDGEMENTS

I want to express my thanks to my family and friends for their encouragement and support to seek educational pursuits throughout my life. I also owe a great deal to Matthew Wilcox for spending countless hours and effort in collecting data and then being so willing to share it with me for my research.

I am grateful to each member of my committee whose intelligence, experience, and perspectives have brought added insight into my research and into my education. I owe special thanks to Dr. Richard Sudweeks for his direction, scholarship, and commitment to teaching which have in large part made the EIME program what it is.

I am especially grateful to my chair, Dr. Ellie Young, for helping me to become more than I thought I could be and to be a better version of myself. Her unfailing encouragement, insight, and mentoring were invaluable to me throughout this whole process. I am grateful for her time spent reading and thoughtfully editing countless drafts, her kind and generous praise, her empathetic listening ear, her endless hope and optimism, and her compassionate guidance. I sincerely could not have done this without her.

TABLE OF CONTENTS

LIST OF TABLES FOR STUDY 1

LIST OF TABLES FOR STUDY 2

LIST OF FIGURES FOR STUDY 1

LIST OF FIGURES FOR STUDY 2

LIST OF FIGURES FOR APPENDIX C

DESCRIPTION OF DISSERTATION STRUCTURE

This dissertation is a combination of two research studies formatted as separate journal-style articles. Both studies draw on examining the psychometric soundness of the *Student Risk Screening Scale for Internalizing and Externalizing Behaviors* (SRSS-IE) and utilize methods based in Item Response Theory (IRT). The SRSS-IE is a screening instrument that attempts to identify students in schools who may be at risk for emotional and behavioral disorders (EBD). The first study is an examination of how the category response options for each item on the SRSS-IE function within each item for a sample of middle school students. The second study is an examination of systematic differences in this same sample of middle school student in how males and females are rated on the SRSS-IE and how such differences could potentially impact cut scores and decisions made regarding which students are at highest risk for EBD.

Study 1 includes evidence that each item's category response options may not currently function as intended by the original scale developers and that revisions to scale items or scale scoring could improve the psychometric functioning of the SRSS-IE. Study 2 includes evidence that students are rated differentially, at least to some degree, by teachers on some of the items on the SRSS-IE. Additionally, cut score analyses demonstrated that the psychometric information provided by the SRSS-IE differs to some degree by gender. Appendices A, B, and C contain relevant computer software syntax and graphs of all items examined on the SRSS-IE.

Both studies suggest revisions to the current SRSS-IE, or revisions to the scoring and cut scores used, to determine whether or not students in middle schools may be at risk for EBD. The aim of this research is to improve the accuracy with which students are classified as being at risk for EBD with the hope that such an improvement in accuracy can lead to providing timely and needed interventions and supports for students who could benefit the most.

ABSTRACT – Study 1

This research study examined the psychometric properties of the *Student Risk Screening Scale for Internalizing and Externalizing Behaviors* (SRSS-IE; Lane, Oakes, et al., 2012) using Item Response Theory (IRT) methods among a sample of 2,122 middle school students. The SRSS-IE is a recently revised screening instrument aimed at identifying students who are potentially at risk for emotional and behavioral disorders (EBD). Utilizing the Nominal Response and Generalized Partial Credit models of IRT, items from the SRSS-IE were evaluated in terms of the degree to which the response options for each item functioned as intended by the scale developers and how well those response options discriminated among students who exhibited varying levels of EBD risk. Results from this study indicate that the four response option configurations of the items on the SRSS-IE may not adequately discriminate among the frequency of externalizing and internalizing behaviors demonstrated by middle school students. Recommendations for item response option revisions or scale scoring revisions are discussed in this study.

Keywords: Student Risk Screening Scale, emotional and behavioral disorders, universal screening, Item Response Theory, Nominal Response Model

An Examination of the Psychometric Properties of the SRSS-IE Using the Nominal Response and Generalized Partial Credit Models

Introduction

The *Student Risk Screening Scale for Internalizing and Externalizing Behaviors* (SRSS-IE) is a universal screening instrument used in a variety of school settings to assess students who may be at risk for emotional and behavioral disorders (EBD). The basis of the SRSS-IE is the *Student Risk Screening Scale* (SRSS), which was "initially developed to detect elementary-age students at risk for antisocial behavior patterns" (Lane, Menzies, Oakes, & Kalberg, 2012, p. 94). Such behavior patterns often manifest as externalizing behaviors, or misbehavior directed toward others (e.g., stealing, lying, aggressive behaviors; Lane, Menzies, Oakes, & Kalberg, 2012). In 2012, Lane, Oakes, and colleagues expanded the SRSS by developing items to address internalizing behaviors (e.g., feeling anxious, sad, withdrawn) in identifying students who may be at risk for a broader range of emotional and behavior disorders. The purpose of this study is to examine whether the items on the SRSS-IE are performing as expected according to the current design of the SRSS-IE and to determine the degree of measurement precision of its items using a sample of middle school students.

While some research examining the psychometric properties on this expanded instrument (i.e., the SRSS-IE) has been done (e.g., Lane, Oakes, et al., 2012), Item Response Theory (IRT) has not been used to empirically examine item functioning or the psychometric soundness of the instrument, especially among a sample of middle school students at a developmental time period that may be especially sensitive to developing EBD risk factors (Hardy, Bukowski, & Sippola, 2002; Lane, Parks, Kalberg & Carter, 2007; Siedman, Allen, Aber, Mitchell & Feinman, 1994). An IRT approach to examining this instrument could provide evidence that the items are

functioning as intended and represent the constructs purported to be measured. More importantly, IRT can be used to ensure that as items function as intended, proper scoring procedures can be utilized and trusted to make determinations about whether individual students need additional assistance in schools.

Ultimately, IRT can provide additional evidence to help ensure that the SRSS-IE is precisely measuring the expanded latent constructs that it purports to measure (i.e., externalizing and internalizing concerns). Such evidence could support the efficacy of the instrument which, in turn, could better help identify students who are actually at-risk and in need of additional support in schools. Additionally, an IRT analysis of the SRSS-IE data in this context could provide evidence of the psychometric soundness of the instrument in order to ensure that it is working as intended among middle school students during early adolescence.

Review of Literature

Screening for Emotional and Behavioral Concerns

Universal screening of students to identify the level of risk concerning social, emotional, and behavioral issues in schools is a relatively recent practice but one essential to adopting "prevention-oriented . . . intervention practices" that could readily benefit individual students (Glover & Albers, 2007, p. 118). Screening instruments used on such a widespread basis must prove to be psychometrically sound, among other attributes (Glover & Albers, 2007). Part of being psychometrically sound includes evidence that a given instrument's scores are reliable and can accurately identify specific individuals who could potentially be at risk (Glover & Albers, 2007). Careful examination of screening instruments' psychometric properties, and specifically, well- conducted item analyses, can help ensure that such instruments are indeed technically

sound and are appropriate and helpful in making decisions related to students' level of risk and psychological needs.

Several universal screening instruments have been developed specifically to identify students at risk for EBD. Manifestations of EBD include externalizing behaviors (e.g., aggression, noncompliance) and internalizing behaviors (e.g., social isolation, restricted activity levels; Lane, Oakes, Lusk, Cantwell, & Schatschneider, 2015). These two constructs (i.e., externalizing and internalizing behaviors) have been extensively examined in the research literature (Achenbach & Rescorla, 2001; Hayden & Mash, 2014; Krueger & Markon, 2006; Lilienfield, 2003; McDermott, 1993; McDermott & Weiss, 1995).

For the externalizing construct, researchers have attempted to operationalize and categorize EBD risk into specific, observable behaviors (e.g., Achenbach & Rescorla, 2001). The seven externalizing behaviors or problems on the SRSS-IE include the following: (a) stealing; (b) lying, cheating, sneaking; (c) behavior problems; (d) peer rejection; (e) low academic achievement; (f) negative attitude; and (g) aggressive behaviors (Drummond, 1994; Lane, Oakes et al., 2012). These behaviors can also be thought of as behaviors directed toward others or "undercontrolled" problems (Hayden & Mash, 2014, p. 27).

The internalizing construct may be more difficult to operationalize since behaviors associated with this construct are less observable (Lane, Menzies, Oakes, & Kalberg, 2012). These behaviors are sometimes viewed as "inner-directed" behaviors or "overcontrolled" problems (Hayden & Mash, 2014, p. 27). Lane, Oakes, and colleagues (2012) indicated that some behaviors or problems that ought to be included on the SRSS-IE in the internalizing dimension are (a) emotionally flat; (b) shy, withdrawn; (c) sad, depressed; (d) anxious; and (e) lonely. The inclusion of internalizing behaviors associated with risk of EBD is a crucial

component to developing an instrument that can accurately screen students who may be at risk of EBD.

The externalizing and internalizing constructs are related. One approach to examining child psychopathology has included cluster analyses or identifying "symptom clusters" such as externalizing behaviors and internalizing behaviors as described above (Hayden & Mash, 2014, p. 27; see also McDermott, 1993; McDermott & Weiss, 1995). Given this strong theoretical evidence to support this multidimensional view of these constructs, there is also a relatively high level of comorbidity (e.g., Krueger & Markon, 2006, reported a correlation of $r = .5$) between externalizing and internalizing disorders (Hayden & Mash, 2014; see also Lilienfeld, 2003).

The constructs of externalizing behaviors and internalizing behaviors are of particular interest here because they constitute the operational definition EBD risk and about 20% of youth are estimated to be at risk for some form of EBD (Forness, Freeman, Paparella, Kauffman, & Walker, 2012). That number is particularly striking since "students with EBD struggle to negotiate relationships with teachers and peers and may struggle academically, often impeding their ability to complete high school successfully" (Lane, Oakes, Lusk, et al., 2015, p. 1; see also Wagner et al., 2006). Dropout rates for students with EBD can be as high as 50% and suspension rates for these students around 73% (Wagner, Kutash, Duchnowski, Epstein, & Sumi, 2005). Additionally, students with EBD frequently miss school and even when they attend school, they have academic difficulties ranging from poor task completion to lack of content knowledge (Lane, Wehby, & Barton-Arwood, 2005; Young, Caldarella, Richardson, & Young, 2012). Therefore, developing strong psychometric instruments to accurately detect students who are at risk for EBD in schools is of primary importance and can aid in promoting academic engagement (Reddy, Newman, De Thomas, & Chun, 2009). Also, when at-risk students are

accurately identified, educators can design and implement interventions to address the students' risk factors before concerns and behavior escalate and require more resources and time to address.

Screening Instruments

Instruments that have been developed to screen for EBD include the Systematic Screening for Behavior Disorders (SSBD; Walker & Severson, 1992; see also Walker, Severson, & Feil, 2014), the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997), and the Student Risk Screening Scale (SRSS; Drummond, 1994). In one research study, Lane et al. (2009) compared these three screening instruments in terms of ease of administration, sensitivity, and specificity in correctly identifying students at risk for EBD. The researchers concluded that overall, the psychometric properties of the SRSS were quite similar to the gold standard instrument, the SSBD (see Kauffman & Landrum, 2009; Lane et al., 2009), when used for screening elementary students with the exception of when it was used to identify students at risk who exhibited internalizing behavior patterns (Lane et al., 2009). The comparability of the SSBD and the SRSS for older students (i.e., middle school and high school students), however, is still emerging (e.g., Caldarella, Young, Richardson, Young, & Young, 2008; Kalberg, Lane, Driscoll, & Wehby, 2011).

To further address the comparability of the SSBD and the SRSS at the elementary school level, the study by Lane et al., (2009) was replicated in a subsequent research study with similar results (Lane, Kalberg, Lambert, Crnobori, & Bruhn, 2010). Assuming the primary goal in using the specific screening instrument was to identify students with externalizing behavior patterns, the SRSS was quite comparable to the SSBD in terms of internal consistency, sensitivity, and specificity. If the primary goal, however, was to identify students with internalizing behavior

patterns, the SRSS, understandably, fell short in terms of utility (Lane, Kalberg, Lambert, Crnobori, & Bruhn, 2010).

Such research studies have led to the reexamination of the theoretical construct (or possible multiple constructs) underlying the SRSS as well as the development of additional screening instruments aimed at identifying students at risk for EBD who exhibit externalizing and/or internalizing behavior patterns. One example of such instruments is the Student Risk Screening Scale for Internalizing and Externalizing Behaviors (SRSS-IE). The SRSS-IE is an extension of the seven-item SRSS instrument developed by Drummond (1994) that focused solely on externalizing behaviors. The SRSS-IE contributed additional items relating internalizing behaviors aimed at more fully identifying and measuring constructs related to risk for EBD (Lane, Oakes et al., 2012). Implicit in the structure of the SRSS-IE is that the manifestations or indicators of EBD fall under two separate, distinct theoretical constructs: internalizing behaviors and externalizing behaviors (Lane, Oakes, et al., 2012).

While this newer instrument (i.e., the SRSS-IE) has been researched by Lane and colleagues to support its psychometric soundness (see Lane, Oakes et al., 2012; Lane, Menzies, Oakes, Lambert et al., 2012; Lane, Oakes, Lusk, et al., 2015; Lane, Oakes, Carter, Lambert, & Jenkins, 2013), more research remains to be done. Initial support for the theoretical structure of the latent constructs is evident (e.g., Lane, Oakes et al., 2012); however, questions remain as to whether internalizing behaviors and externalizing behaviors are separate and distinct constructs or rather overlapping constructs under a broader construct of EBD risk (Hayden & Mash, 2014; Krueger & Markon, 2006; Lilienfeld, 2003).

Additionally, while the SRSS has been closely examined at the elementary, middle school, and high school levels (e.g., see Lane et al., 2010; Lane et al., 2011; Lane, Bruhn, Eisner,

& Kalberg, 2010), the SRSS-IE has been primarily researched at the elementary level (see Lane, Menzies, Oakes, Lambert et al., 2012; Lane, Oakes et al., 2012; Lane, Oakes, Lusk, et al., 2015) with some exceptions (e.g., Lane et al., 2013). Given that some research indicates that the median age of onset for anxiety and impulse-control disorders is around age 11 and "half of all lifetime cases start by age 14 years," there is currently a lack of evidence to establish the psychometric soundness of this instrument where it may be critically valuable at the middle school and high school levels (Kessler et al., 2005, p. 593).

An Item Response Theory Approach to Scale Examination

Newly established sophisticated psychometric approaches and methodologies that are useful in examining latent variables could also be helpful in further exploring how well the SRSS-IE is accurately measuring what it purports to measure. One specific methodology that can provide additional evidence of a psychometrically sound instrument is Item Response Theory (IRT).

IRT is a measurement perspective that "is, in effect, a system of models that defines one way of establishing the correspondence between latent variables and their manifestations" (de Ayala, 2009, p. 4). IRT models can be used to locate persons and items along the same continuum which allows researchers to characterize individuals according to their locations on the latent trait continuum and items in terms of their location along that same latent trait continuum (de Ayala, 2009).

IRT has statistical and measurement advantages over Classical Test Theory (CTT) that include: (a) item parameter estimates that are sample-independent; and, (b) importantly, the ability to determine how well a given model fits the data (de Ayala, 2009).

Although Confirmatory Factor Analysis (CFA) is a more common approach to scale construction and development, IRT can also be useful in the process (Bartholomew, Knott, & Moustaki, 2011) and can provide additional insight into item functioning. Item and test information functions produced by IRT analysis methods can also help to determine the psychometric information or precision of measurement provided by each individual item or the test as a whole. Lastly, information from IRT analyses help to determine which item response options are more likely to be endorsed at specific trait levels (de Ayala, 2009).

Category Response Functions

Analysis of category response functions across the latent trait scale using IRT is of particular interest in the current research study. First, exploring the item information functions and category option functions along the latent trait continuum could help provide psychometric evidence of the precision of the instrument in terms of its ability to correctly identify students at a particular trait level indicating risk of EBD. Additionally, in the case of the SRSS-IE, an IRT analysis of item information and category response functioning could help to ensure that the instrument is measuring the full breadth of the externalizing and internalizing constructs.

IRT includes several types of models that can be used to analyze data. The Nominal Response Model (NRM) developed by Bock (1972) was initially created to analyze polytomous items with unordered categories (de Ayala, 2009). However, recent research has demonstrated its usefulness in polytomous models for ordered categories because of the NRM's ability to calculate within and between item differences in the discrimination parameter for each item option (see Preston & Reise, 2015; Preston, Reise, Cai, & Hays, 2011).

Additionally, using the NRM for analyzing items with ordered category options has the advantage of allowing the use of category response option discrimination parameters to calculate

Category Boundary Discrimination (CBD) parameters and category intersection parameters between adjacent category response options within an item (Preston & Reise, 2015; Preston, et al., 2011.) CBD parameters "index the degree to which a particular dichotomous distinction (e.g., a response in category two vs. one) discriminates trait levels," (Preston, et al., 2011, p. 523) whereas category intersection parameters indicate the trait level at which raters are equally likely to endorse adjacent category response options. Both CBD parameters and category intersection parameters are useful in providing evidence as to whether category response options are functioning as empirically ordered categories. If the CBD parameters are all positive values and the category intersection parameters increase for each successive category intersection point, there is evidence that category response options are functioning as ordered categories (Preston & Reise, 2015). Previous research using Rasch IRT models has evaluated category ordering using threshold discriminations and successively increasing step calibrations (Andrich, 1978; Linacre, 2002). Using the NRM to evaluate ordered categories, however, provides statistical evidence as to whether each category response option is accurately ordered in the manner intended and whether teachers using the rating scale can successfully differentiate between adjacent category options.

Using CBD parameters to examine the psychometric properties of a scale can also provide useful information relating to how the instrument is scored. Preston & Reise (2015) indicate that when any of the CBD parameters from the scale are zero (or close to zero), "any scoring strategy…will add nuisance variation into the estimate," (p. 396). In 2011, Preston and colleagues warned that such variation in scoring can distort the scaling of individual differences and, in the case of the SRSS-IE, incorrectly identify students as either at-risk for EBD when they

are not, or, more seriously, fail to identify students as at-risk for EBD when they are (Severson, Walker, Hope-Doolittle, Kratochwill, & Gresham, 2007).

Another IRT model that is useful in analyzing polytomous data is the Generalized Partial Credit Model (GPCM; Muraki, 1997). The GPCM is an IRT model that is nested within the NRM and provides a single discrimination parameter estimate for each item as a whole. Since the GPCM is nested within the NRM, the relative fit of these models can be compared using -2 log likelihood statistics. In addition, they can be compared using the Akaike Information Criterion (AIC; Akaike, 1974) and the Bayesian Information Criterion (BIC; Schwarz, 1978) for non-nested models.

In summary, the SRSS-IE has shown promise in being an important tool in screening children in schools who may be at risk for EBD. Further research, however, is critical in validating this instrument. Such research should include the use of current psychometric techniques to get a more accurate representation of how individual items are performing and if they are performing as intended. Ultimately, the goal of research in attempting to provide empirical evidence for the utility of SRSS-IE is to make the instrument psychometrically sound and more widely available in the field for the benefit of teachers and students.

Research Questions

The following research questions will be addressed in this study:

1. To what degree do each of the four response options on the items of the SRSS-IE function as intended by the scale developers?
2. Which response options, if any, are rarely used and need to be revised or deleted?

3. To what extent are the response options ordered so that the next higher option in each pair of adjacent categories represents an increase in frequency of occurrence of the specified behavior?
4. To what extent do the response options discriminate among students who exhibit varying degrees of externalizing and internalizing behaviors?

Method

Participants and Setting

In order to help answer these research questions, de-identified, archival data from 2,122 students from local middle schools were analyzed. The original ratings were generated at the middle schools from 93 teachers (61% female) who completed the SRSS-IE for each student in their first-period (homeroom) class.

The existing data were initially gathered from a large, suburban school district in the mountain west. The data came from three middle schools within the district and represent information about students in grades 6 through 8. All available data were used for this study. Grade and gender demographic information for the sample is displayed in Table 1. Table 2 contains demographic information about racial backgrounds for both teachers and students in this research study:

Table 1

Frequency Distribution of Students' Gender by Grade

	Male		Female		
Grade	n	%	n	%	Total
6	372	47%	420	53%	792
7	364	54%	311	46%	675
8	346	53%	309	47%	655
Total	1082	51%	1040	49%	2122

Table 2

Frequency Distribution of Teachers' and Students' Racial Backgrounds

	Teachers (n = 93)		Students (n = 2,122)	
Race	n	%	n	%
White	64	69%	1664	78%
Hispanic/Latino	16	17%	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%

Measure

The SRSS-IE was developed by Lane, Oakes et al. (2012). The SRSS-IE consists of 12 items (seven items representing the externalizing construct and five items representing the internalizing construct) in which a teacher rates each student on the frequency of each observed behavior. The frequency scale has four response options ranging from 0 (*never*), 1 (*occasionally*), 2 (*sometimes)*, to 3 (*frequently*). The seven items representing externalizing behaviors are (a) stealing; (b) lying, cheating, sneaking; (c) behavior problems; (d) peer rejection; (e) low academic achievement; (f) negative attitude; and (g) aggressive behaviors. The five items representing internalizing behaviors are (a) emotionally flat; (b) shy, withdrawn; (c) sad, depressed; (d) anxious; and (e) lonely. The teachers rated each student on each of these 12 behaviors.

Several studies have been conducted to demonstrate the reliability of scores and validity evidence of this instrument. Internal consistency reliability estimates for the SRSS-IE items at the elementary level were initially shown to be $\alpha = .84$ for the externalizing behavior items and $\alpha = .72$ for the internalizing behavior items (Lane, Oakes et al., 2012). Examination of the factor structure of the data using an exploratory factor analysis (EFA) at the elementary level also

revealed a strong two-factor model (i.e., externalizing and internalizing) with the peer rejection item loading fairly high on both latent constructs (Lane, Oakes et al., 2012). Evidence of convergent validity of the SRSS-IE with previously existing instruments (e.g., SDQ) has been demonstrated and correlations of the total scores on the SRSS-IE to the SDQ subscales has ranged from .49 to .75 for students at the elementary school level (Lane, Oakes et al., 2012). Subsequent studies of the SRSS-IE have demonstrated similar reliability and validity results at the elementary school level (e.g., see Lane et al., 2010; Lane, Bruhn, Eisner, & Kalberg, 2010), as well as the middle school level (Lane, et al., 2013), and at the high school level (Lane et al., 2011; Lane, Oakes, Cantwell, Menzies, et al., 2016).

The SRSS-IE data analyzed in this study were collected previously from the teachers and the psychometric analyses were done on the existing data set. Demographic information about the students' gender and race, and the teachers' demographic characteristics were collected along with the SRSS-IE data (Wilcox, 2016).

Subscale scores on the SRSS-IE are calculated by adding scores for each of the seven externalizing items for each student to produce an externalizing risk score ranging from 0 to 21. The scores for each of the five internalizing items for each student are then added together to yield an internalizing risk score ranging from 0 to 15 (Lane, Oakes, Common, et al., 2015). For the externalizing subscale, students are classified according to the following risk categories: low risk (0-3); moderate risk (4-8); and, high risk (9-21; Lane, Menzies, Oakes, & Kalberg, 2012). For the internalizing subscale, cut scores to determine low, moderate, and high risk for students were not initially established (Lane, Oakes, Lusk, et al., 2015); however, recent studies at the elementary level have proposed that the categories should be: low risk (0-1); moderate risk (2-3); and, high risk (4-15; Lane, Oakes, Swogger, et al., 2015). Research about possible cut scores for

use in classifying at-risk students at the middle school and high school levels is also being investigated (Lane, Oakes, Cantwell, Schatschneider, et al., 2016).

Design and Analyses

To examine the research questions in this study, IRT methodologies were used. Since IRT is based on the assumption that the items being analyzed measure a unidimensional construct, the data for each of the SRSS-IE factors were analyzed separately. The analyses were done using the flexMIRT software (Cai, 2013) followed by the use of R 3.3.1 (R Core Team, 2014) and RStudio 0.99.903 (RStudio Team, 2015). Two different IRT models were used for these data. First, the Nominal Response Model (NRM) was used (Bock, 1972, 1997) to assess category response functioning. This was done by examining category discrimination parameters and calculating CBD parameters by subtracting adjacent category discrimination parameters. Low or negative CBD parameter estimates indicate category options providing little or no information while high CBD parameter estimates are highly informative (Preston & Reise, 2015). Category intersection points were also calculated using the results of the CBD parameter estimations.

Next, the Generalized Partial Credit Model (GPCM) was used (see Muraki, 1997) and item discrimination parameters were generated using the GPCM and CBD parameters were produced using the NRM. Model fit of the data using NRM and GPCM was compared to determine which model best fit these data. Comparisons with the NRM helped to determine whether the more parsimonious GPCM could be used without loss of psychometric information (Preston, et al., 2011; Preston, 2014a). Decisions about the fit of each model were made based on the -2 log likelihood statistics, the AICs, and the BICs for each model from the flexMIRT output.

Parameter estimates were calculated in flexMIRT first using the NRM and then using the GPCM for each construct in the SRSS-IE (see Appendix A for sample flexMIRT code). The flexMIRT output was then used in RStudio (see Appendix B for sample RStudio code) to create Category Response Curves (CRCs) and item and category information plots (Preston, 2014b). To examine the degree to which each of the four response options functioned as intended and whether all response options were used, the CRCs for each item were carefully examined. Special attention was paid to category response thresholds in order to detect any category response curves which entirely overlapped another.

CBD parameters were then calculated followed by a Wald test which was used to determine whether within-item CBD parameters varied significantly (Preston 2014c; Wald, 1945). Item options with statistically significant ($\alpha < .05$) CBD parameters were collapsed with adjacent category options and item analyses were re-run in flexMIRT using the new category option configurations. This process was repeated until the Wald test statistics were no longer significant or until the items became dichotomous.

Item information functions were also examined to help answer how well the range of item difficulties adequately represents the whole range of the latent variables of interest (i.e., externalizing and internalizing behaviors). Item information functions should spread over a wide range of trait levels and should peak at specified trait levels ensuring that the items on the SRSS-IE are sufficiently covering the breadth of the constructs.

Results

Each analysis described in the method section was performed and all model estimation procedures terminated normally (i.e., convergence criteria were satisfied in the IRT models). The following are the results from the IRT analyses of the 12 SRSS-IE items.

Comparison of Results Using the NRM and GPCM

To determine how the four response options of each item functioned and to examine the order and discrimination properties of the options for each item, the estimated item discrimination parameters were retrieved from the flexMIRT output. Subsequently, CBD parameters were computed and are displayed with the discrimination parameters using the GPCM in Tables 3, 4, 5, and 6.

Table 3

Discrimination Parameters Produced by the NRM and GPCM for the Externalizing Items

		NRM			GPCM
Item	Content	a_{1*}	a_{2*}	a_{3*}	a_{GPCM}
1	Stealing	2.72	1.51	0.60	2.54
2	Lying, cheating, sneaking	2.87	2.10	1.29	2.39
3	Behavior problems	2.25	1.15	2.04	1.98
4	Peer rejection	1.25	0.51	0.49	0.96
5	Low academic achievement	1.33	0.82	0.52	0.96
6	Negative attitude	1.67	1.59	0.90	1.59
7	Aggressive behaviors	1.99	1.11	1.27	1.75
Mean		2.01	1.26	1.02	1.74

Table 4

Model Fit of Externalizing Items

	Model		
	NRM	GPCM	Difference
-2 Log likelihood	17811.97	17882.01	70.01
Akaike Information Criterion (AIC)	17895.97	17938.01	42.04
Bayesian Information Criterion (BIC)	18133.69	18096.49	-37.20

Table 5

Discrimination Parameters Produced by the NRM and GPCM for the Internalizing Items

		NRM			GPCM
Item	Content	a_{1*}	a_{2*}	a_{3*}	a_{GPCM}
8	Emotionally flat	1.94	0.73	1.40	1.56
9	Shy, withdrawn	1.34	0.53	0.43	0.85
10	Sad, depressed	3.01	1.84	1.36	2.66
11	Anxious	1.30	0.36	0.88	0.98
12	Lonely	3.24	2.33	3.76	3.02
Mean		2.17	1.16	1.57	1.81

Table 6

Model Fit of Internalizing Items

	Model		
	NRM	GPCM	Difference
-2 Log likelihood	12706.26	12770.38	64.12
Akaike Information Criterion (AIC)	12766.26	12810.38	44.12
Bayesian Information Criterion (BIC)	12936.07	12923.58	-12.49

Category intersection parameters were also calculated for all items on the SRSS-IE and are displayed in Table 7.

Table 7

Category Intersection Parameters Produced by the NRM for All Items on the SRSS-IE

Item	Intersection 1	Intersection 2	Intersection 3
Stealing	1.765	2.636	4.233
Lying, cheating, sneaking	0.951	1.338	1.992
Behavior problems	1.164	1.191	1.681
Peer rejection	1.848	2.490	3.449
Low academic achievement	0.947	0.951	0.885
Negative attitude	1.281	1.340	1.889
Aggressive behaviors	1.563	2.063	2.197
Emotionally flat	1.407	1.877	1.957
Shy, withdrawn	1.231	1.377	1.349
Sad, depressed	1.176	1.701	2.404
Anxious	1.992	2.972	2.216
Lonely	1.188	1.584	2.122

All 12 items of the SRSS-IE demonstrated significant CBD parameter differences according to the Wald tests ($\alpha < .05$). Category intersection parameters also revealed a concern with the "low academic achievement," "shy, withdrawn" and "anxious" items since the third intersection point was smaller than the second intersection point for each of the items, suggesting the upper categories on these items may be unordered. All other category intersection parameters on the remaining nine items increased at each intersection point, suggesting that the categories are functioning as ordered categories on these items.

Following these analyses, revisions to the category response options on each of the items were performed. Since larger CBD parameters indicate more informative category option functioning, the smallest CBDs were identified and subsequently, adjacent category response options were combined in the next iteration of item and category parameter estimation (Preston & Reise, 2015). This process was repeated twice with each item that continued to have significantly different CBD parameter estimates. The results of this process were category response option configurations that were empirically ordered and that improved item and category information.

For the externalizing construct, items 1, 4, 5, and 6 (stealing; peer rejection; low academic achievement; and negative attitude, respectively) had significantly different CBD parameters throughout each iteration of the item revision process. According to this model, these items were best depicted as dichotomous items rather than polytomous items. For these four items, item options 1, 2, and 3 (i.e., "occasionally," "sometimes," and "frequently") were combined leaving the items consisting of option 0 ("never") and a combination of some frequency of the externalizing behaviors. Items 2, 3, and 7 (lie, cheat, sneak; behavior problems; and aggressive behaviors, respectively) had significantly different CBD parameters in only the

first round of parameter estimation. Categories with the smallest CBD parameters were collapsed into adjacent categories in the subsequent analysis. For item 2, options 2 and 3 (i.e., "sometimes" and "frequently") were combined; for items 3 and 7, options 1 and 2 (i.e., "occasionally" and "sometimes") were combined. After estimation using this new three-option configuration, CBD parameter differences were no longer significant, indicating that the three-option items were a better depiction of the item options than the original four-option items.

Illustrative examples of visual depictions of individual item CRCs are displayed in Figures 1 and 3 while their respective item and category information curves for two of the seven externalizing items are displayed in Figure 2 and 4. (See Appendix C for all 12 item CRCs with their respective item and category information functions.)

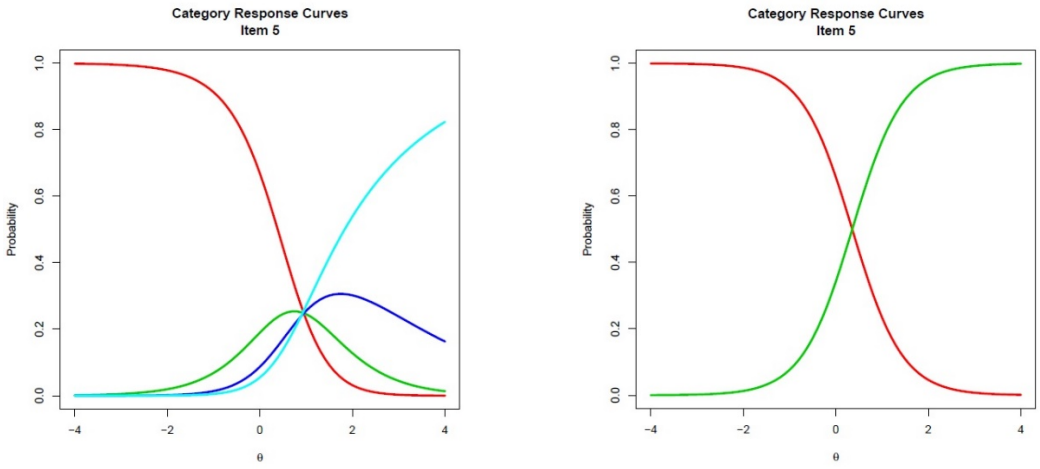

Figure 1. Initial (left) and final (right) CRCs for Item 5, low academic achievement, on the SRSS-IE.

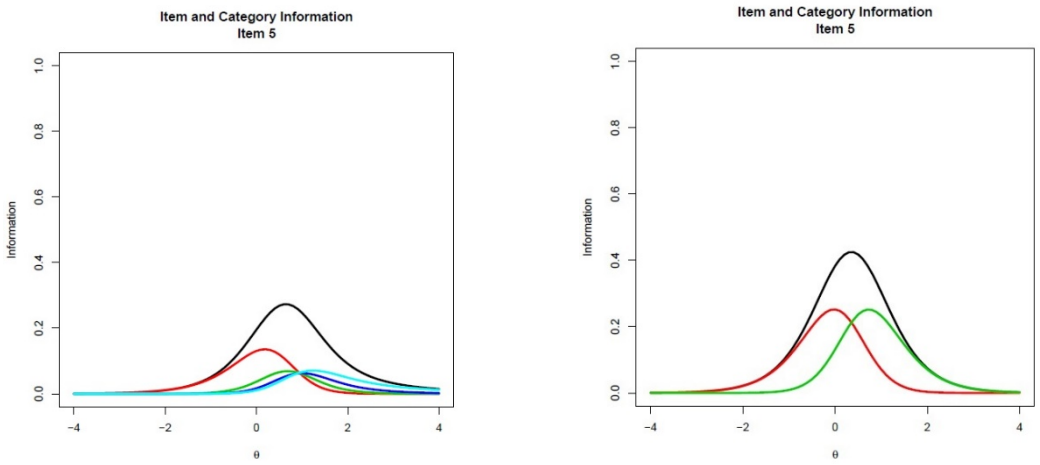

Figure 2. Initial (left) and final (right) item and category information functions for Item 5, low academic achievement, on the SRSS-IE.

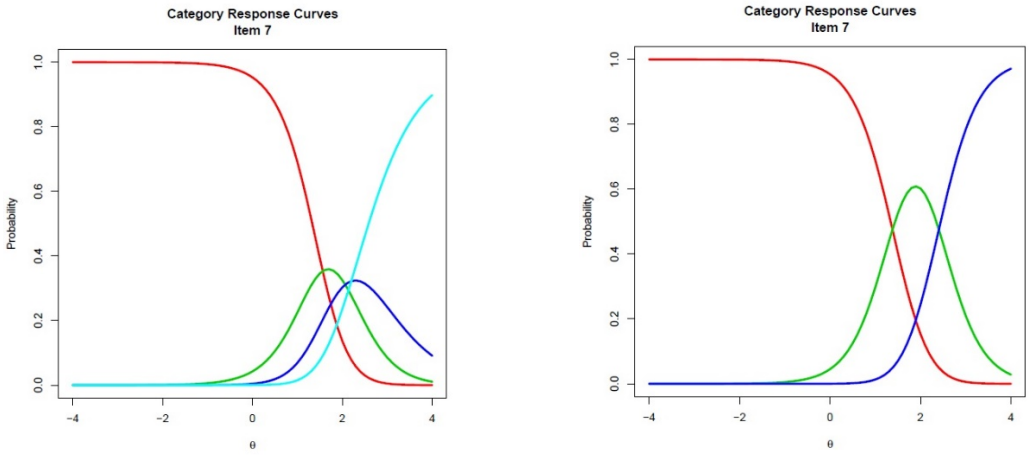

Figure 3. Initial (left) and final (right) CRCs for Item 7, aggressive behaviors, on the SRSS-IE.

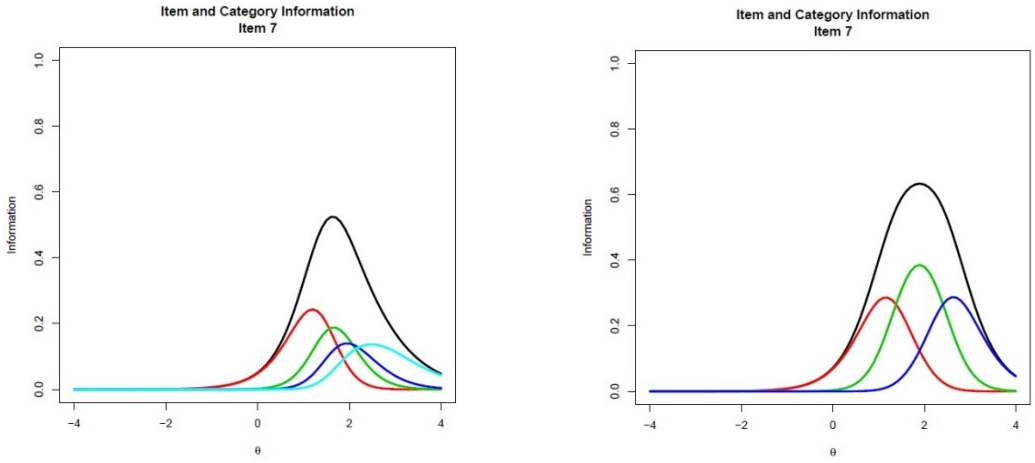

Figure 4. Initial (left) and final (right) item and category information functions for Item 7, aggressive behaviors, on the SRSS-IE.

These two externalizing items depict changes in CRCs and item and category information functioning when item options were collapsed. The original item 5, "low academic achievement," demonstrated significant overlapping of CRCs and little information provided by each category option. The revised item 5, however, demonstrated a clear distinction between only two category response options and each response option provided more psychometric information at both the item and category levels. The original item 7, "aggressive behaviors," also demonstrated significant overlapping of CRCs and little information provided by each category option. The revised item 7, however, demonstrated a clear distinction between three category response options and each response option subsequently provided more psychometric information at both the item and category levels.

For the internalizing construct, items 9 and 10 (shy, withdrawn; and sad, depressed, respectively) had significantly different CBD parameters throughout each iteration of the item revision process. According to this model, these items were best depicted as dichotomous items rather than polytomous items. For these two items, item options 1, 2, and 3 (i.e., "occasionally,"

"sometimes," and "frequently") were combined leaving the items consisting of option 0 ("never") and a combination of some frequency of the internalizing behaviors. Items 8, 11, and 12 (emotionally flat; anxious; and lonely, respectively) had significantly different CBD parameters in only the first round of parameter estimation. Categories with the smallest CBD parameters were collapsed into adjacent categories in the subsequent analysis. For items 8, 11, and 12, options 1 and 2 (i.e., "occasionally" and "sometimes") were combined. After estimation using this new three-option configuration, CBD parameter differences were no longer significant, indicating that the three-option items were a better depiction of the item options than the original four-option items.

Visual depictions of individual item CRCs are displayed in Figures 5 and 7 while their respective item and category information curves for two of the five internalizing items are displayed in Figure 6 and 8.

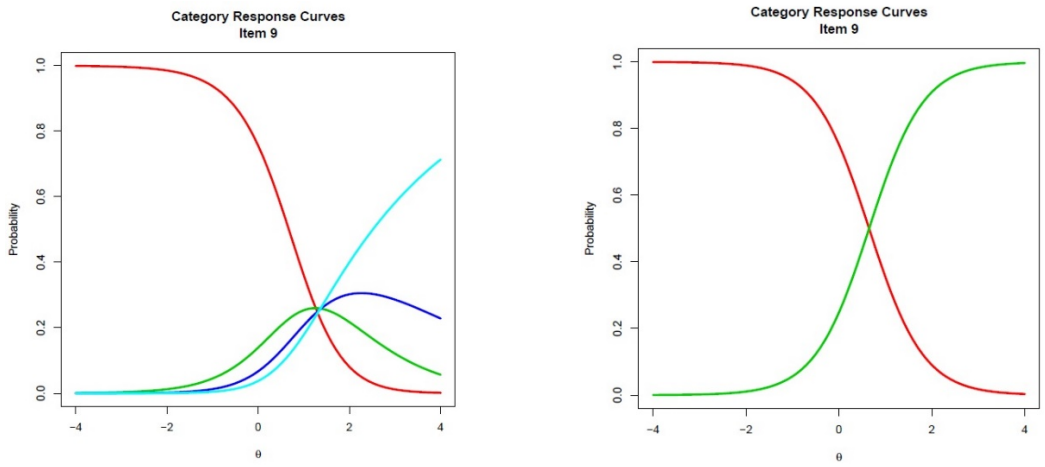

Figure 5. Initial (left) and final (right) CRCs for Item 9, shy/withdrawn, on the SRSS-IE.

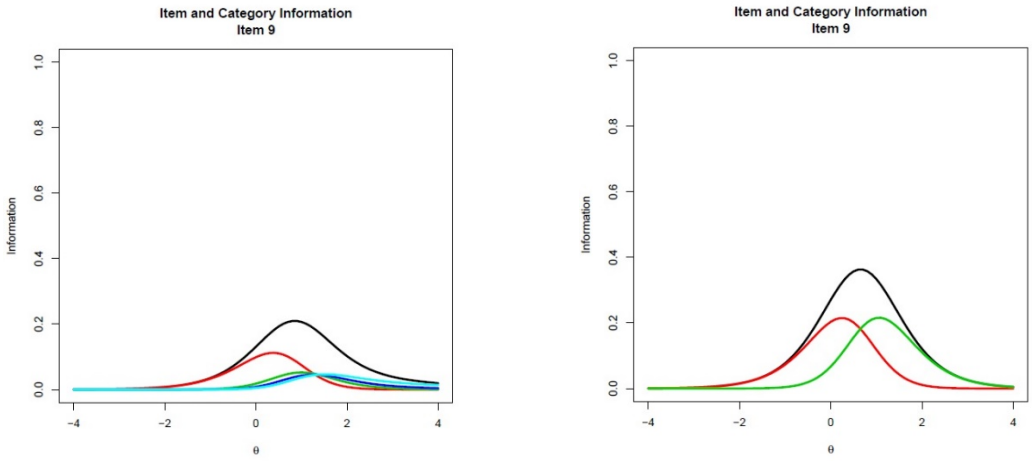

Figure 6. Initial (left) and final (right) item and category information functions for Item 9, shy/withdrawn, on the SRSS-IE.

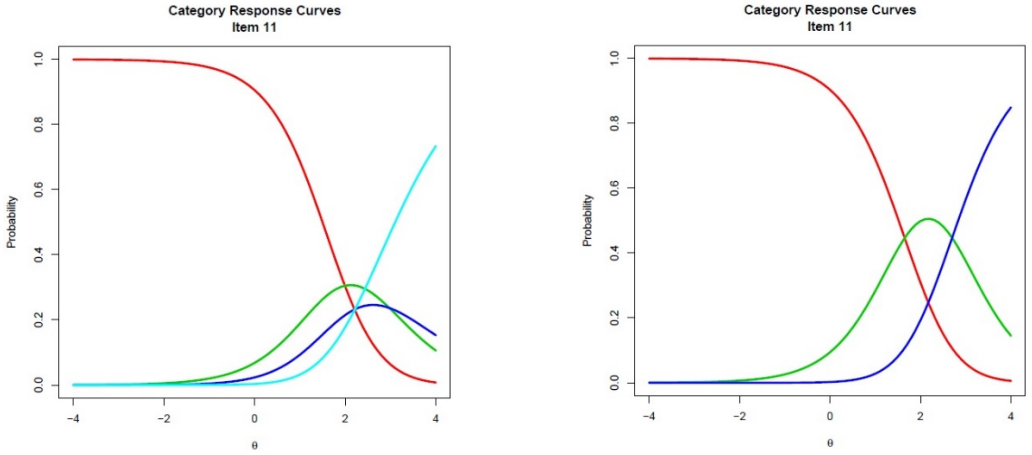

Figure 7. Initial (left) and final (right) CRCs for Item 11, anxious, on the SRSS-IE.

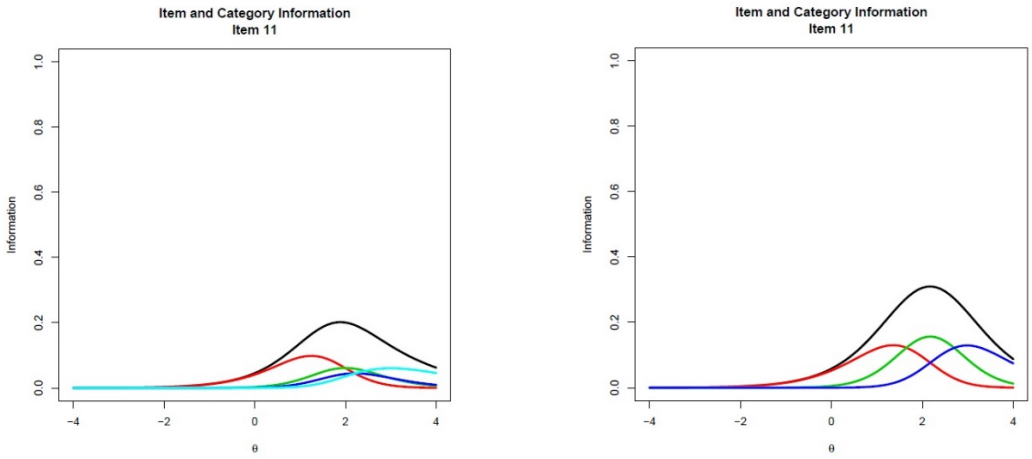

Figure 8. Initial (left) and final (right) item and category information functions for Item 11, anxious, on the SRSS-IE.

These two internalizing items depict changes in CRCs and item and category information functioning when item options are collapsed. The original item 9, "shy, withdrawn," demonstrates significant overlapping of CRCs and little information provided by each category option. The revised item 9, however, demonstrates a clear distinction between only two category response options and each response option provides more psychometric information at both the item and category levels. The original item 11, "anxious," also demonstrates significant overlapping of CRCs and little information provided by each category option. It should also be noted here that the original item exhibited unordered category response options. The revised item 11, however, demonstrates a clear distinction between three, ordered category response options and each response option now provides more psychometric information at both the item and category levels.

The test information function for the original externalizing scale is shown on the left side of Figure 9. In contrast, the test information function for the revised externalizing scale is shown on the right.

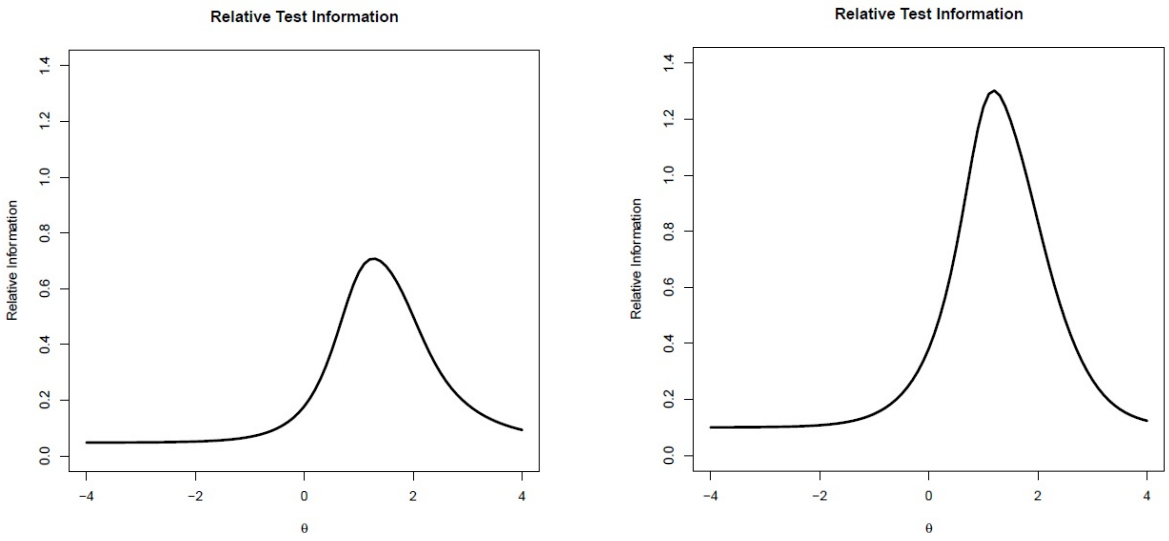

Figure 9. Test information functions for the externalizing scale of the SRSS-IE.

The limited-information fit statistics of the fitted model from the flexMIRT output are also included for reference in Table 8.

Table 8

Limited-Information Fit Statistics of the Fitted Model for the Externalizing Items

Type of Options	M2	df	p-value	F_0hat	RMSEA
Original	612.40	168	0.0001	0.2886	0.04
Revised	183.34	35	0.0001	0.0864	0.04

Note. The Tucker-Lewis fit index based on M2 was 0.95 for the original items and 0.98 for the revised items.

The test information function for the original internalizing scale is shown on the left side of Figure 10. In contrast, the test information function for the revised internalizing scale is shown on the right.

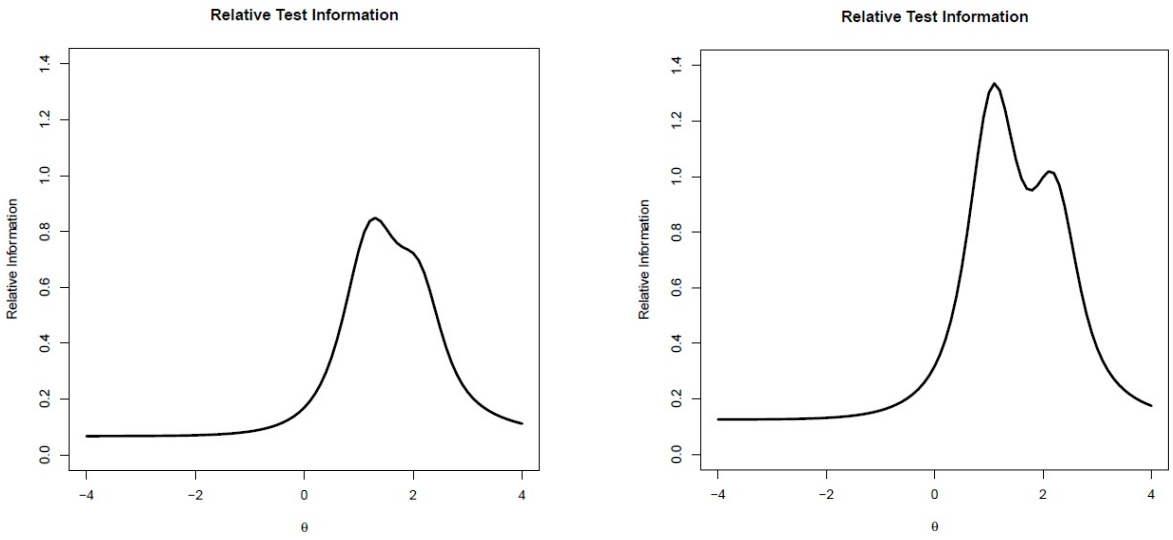

Figure 10. Test information functions for the internalizing scale of the SRSS-IE.

The limited-information fit statistics of the fitted model from the flexMIRT output are also included for reference in Table 9.

Table 9

Limited-Information Fit Statistics of the Fitted Model for the Internalizing Items

Type of Options	M2	df	p-value	F_0hat	RMSEA
Original	368.04	75	0.0001	0.1734	0.04
Revised	128.13	20	0.0001	0.0604	0.05

Note. The Tucker-Lewis fit index based on M2 was 0.94 for the original items and 0.97 for the revised items.

Post hoc descriptive statistics were done using SPSS (version 24.0) to examine the impact of combining category response options as described in this research study on the subscales' total scores. The mean and variance of the total scores on both subscales of the SRSS-IE were compared both before and after revisions. The amount of variance decreased for both the externalizing and internalizing subscales in the revised versions of the items on the SRSS-IE. Results are displayed in Table 10.

Table 10

Mean and Variance of the Total Item Scores for Both Subscales of the SRSS-IE

	Original Items		Revised Items	
Scale	Mean	Variance	Mean	Variance
Externalizing	2.54	13.43	1.70	5.04
Internalizing	1.66	6.98	1.09	2.51

The relationship between original and revised total scores for each subscale were also examined using correlation coefficients with results of r = .965 for the externalizing subscale and r = .958 for the internalizing subscale. Strong, positive, linear relationships between the original scale scores and the revised scale scores are depicted in the scatterplots of these correlations in Figure 11.

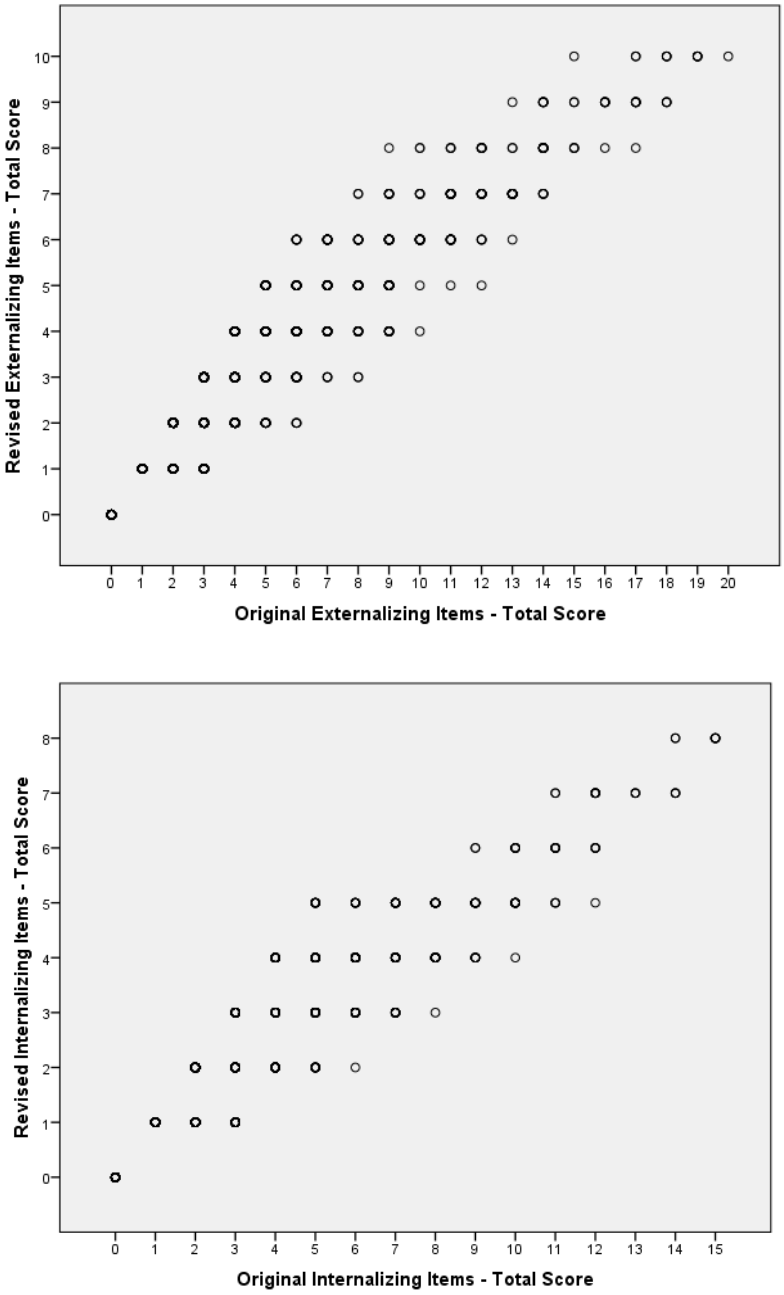

Figure 11. Scatterplots of total scores of the original and revised versions of the items on the externalizing subscale (top) and the internalizing subscale (bottom).

Discussion

Item Response Theory (IRT) methods can be useful in providing psychometric evidence for the precision of measurement of specific constructs used in screening instruments such as the SRSS-IE. IRT analyses contribute to understanding the construct being measured and how such

measurements can be used to distinguish among middle school students who might be more at risk for EBD than others according to teacher perceptions.

Using the NRM to assess item functioning for polytomous items is particularly helpful in determining the degree to which each of the four response options on the items of the SRSS-IE are functioning as intended, which can influence the accuracy of scoring procedures and subsequent decisions made based on these scale scores (Preston & Reise, 2015). It is also useful in determining if category response options could be revised or deleted and to what extent the response options discriminate among individuals.

Results from this study suggest that the category response options presented with the SRSS-IE items did not help the teachers make meaningful distinctions among the frequency of the externalizing and internalizing behaviors for middle school students. Possible revisions or deletions of category response options could greatly improve each item's ability to discriminate among individuals and provide the greatest amount of measurement information available to aid in more meaningful interpretation of the screening results and understanding how the construct is expressed during the middle school years (Preston & Reise, 2015). Another option to improve the measurement precision of the SRSS-IE without changing current item configurations could include collapsing category response options as described in this study only when scoring the items in order to more accurately determine the level at which teachers perceive that students exhibit externalizing and internalizing behaviors (Preston, 2014a).

Examination of the CRCs and item and category information curves reveal that the original four-response option configuration of items on the SRSS-IE may not accurately represent teachers' abilities to discriminate meaningfully among the four category response options (i.e., never, occasionally, sometimes, frequently). All 12 items on the SRSS-IE show

some overlapping CRCs which indicate that at certain levels of the trait (i.e., levels of exhibiting externalizing or internalizing behaviors), raters are equally likely to endorse adjacent category response options for students at the same level of the trait. In other words, the present category response options do not accurately discriminate among individuals with varying levels of EBD risk.

Functioning of the Response Options

While earlier studies indicate that elementary teachers seem to be able to rate the behavior of students on a continuum (see Lane, Menzies, Oakes, Lambert, et al., 2012), these results indicate that middle school teachers who rate their students on the SRSS-IE may be only able to make a meaningful distinction between whether a student does not manifest some targeted behaviors at all (i.e., indicated by "never" on the SRSS-IE) and whether that student does a behavior some of the time (i.e., indicated by "occasionally," "sometimes," or "frequently" on the SRSS-IE) for the stealing, peer rejection, low academic achievement, negative attitude, shy/withdrawn, and sad/depressed items. These results suggest that a revision of such items should consider providing only dichotomous response options for these items on the SRSS-IE. In doing so, the amount of information and the precision of measurement of these items would greatly increase. This change would also likely result in category and item information curves that are taller and are peaked at critical levels of theta that are used in determining whether or not a student might be at risk for EBD.

A possible explanation for differing amounts of overlap in the CRCs is that specific behaviors may occur over a more restricted continuum than other behaviors addressed in specific items or that teachers assume that behaviors are either present or not present, so that the manifestation of certain behaviors seems more discrete. These results may also indicate that

teachers perceive that students at this developmental stage do not display the behavior over a continuum, but rather as present (indicating some risk of EBD) or not present. Examples of these discrete patterns of behavior are seen in the items that address stealing, peer rejection, low academic achievement, negative attitude, shy/withdrawn, and sad/depressed (McDermott, 1993).

Ordering of Response Options

Using the NRM provides an empirical method to determine whether category response options are ordered according to the original design of the instruction (i.e., in the case of the SRSS-IE, increasing frequency of the behaviors addressed; Samejima, 1996). Positive CBD parameters as well as category intersection parameters that sequentially increase all provide evidence that category response options are ordered. Although all CBD parameters in the SRSS-IE were positive, one item ("anxious") exhibited category intersection parameters that were not ordered at higher levels of the trait. The "low academic achievement" item, although category response options were technically ordered, also produced category intersection parameters that were nearly identical at the low end of the trait, suggesting that category response options may not be ordered as clearly as the test authors may have designed.

Item and category information should be maximized by combining some adjacent category response options in order to increase the psychometric measurement precision for the SRSS-IE as a screening tool that can make meaningful distinctions between those who might be at greater risk for EBD than others (see Murray, Booth, & Molenaar, 2016). The "low academic achievement" item provides a good example of both category information functions and overall item information functions increasing even though the category response options are reduced to from four to only two. The above graphical depictions of the category and item information functions all peak at higher levels and, therefore, provide more information, or measurement

precision, than do the category and item information functions of the original four-option "low academic achievement" item. This type of reduction in the number of category response options increased category and item information for all 12 items on the SRSS-IE and the category options then functioned as properly ordered categories.

Additionally, there is evidence that middle school teachers in this sample could only make meaningful distinctions among three category options (as opposed to four) for the items relating to lying/cheating/sneaking, behavior problems, aggressive behaviors, flat emotion, anxiety, and loneliness. These results suggest revising these items to be represented with three category options instead of four. The Wald tests conducted on the CBD parameters for these items were significant when four category options were modeled in the analyses. However, for these items when the Wald tests were repeated with the three category option configuration (i.e., two adjacent categories were combined), the CBD parameters were no longer significantly different, meaning that the categories were each contributing unique information about the construct. Again, a revision of these items with a three category response option pattern would increase the amount of measurement information and measurement precision of the SRSS-IE and demonstrate empirical evidence of properly ordered response options.

Item and Category Discrimination

The GPCM used in the data analyses provided a single overall item discrimination parameter for each polytomous item. Analyses using NRM, on the other hand, provided discrimination parameters for each adjacent pair of category options within each item. The additional psychometric information provided by using the NRM gave additional insight into the polytomous item functioning of each item and category response option on the SRSS-IE. This

allowed for recommended changes to not only items, but to individual response options to the items.

If item and category response revisions are made to all of the items as suggested above, the relative item information and overall test information for the SRSS-IE as a whole would likely increase (Preston & Reise, 2015). This increase in measurement precision is related to an increase in the item and category discrimination parameters as compared with the original items on the SRSS-IE. When this change was made, each item and the screening instrument as a whole, increased in measurement information about students' behavior and increased measurement precision in distinguishing between those students who may be at higher risk of EBD than others.

Scoring Implications of Changes to the SRSS-IE

The scoring methods and cut scores determining levels of EBD risk on the SRSS-IE would need to be modified if these recommendations for item revisions are utilized (Lane, Oakes, Swogger, et al., 2015). However, such modifications could possibly simplify the rating and scoring process for teachers since fewer item options need to be considered; however, cut scores would also need to be adjusted to reflect new scoring procedures. The current system of using cut scores suggested by Lane, Oakes, Swogger, and colleagues (2015) could be used to indicate low, moderate, and high levels of risk of EBD according to total scores on the externalizing and internalizing constructs. However, revisions to the instrument suggested here could provide increased measurement precision that would improve confidence in cut points and possibly reduce the number of cut scores needed to accurately screen students in need of additional support.

New item configurations, however, may necessitate changes in scoring that could include items being grouped according to number of response options (i.e., two or three). Additionally, multiple scoring procedures might be needed in order to determine levels of EBD risk. For example, scoring options could include varying item weights according to the amount of psychometric information provided by that item or the discriminative abilities of that item. Cut scores may also vary according to the number of category response options provided by each item. Most importantly, scoring procedures to assess levels of EBD risk need to be carefully considered to ensure sufficient sensitivity and specificity in correctly identifying students most in need of additional support.

Problematic Item Functioning

Lastly, some of the items offer more overall measurement information about the externalizing and internalizing constructs than others. For example, examination of the item information curves reveals that "behavior problems" is much more informative than "peer rejection" at a given level of theta on the externalizing construct and "emotionally flat" is much more informative than "anxious" at a given level of theta on the internalizing construct. Items with higher peaks on the item information curves contribute more to the overall test information, and therefore measurement precision, than do items with lower peaks. With the revised items and scales suggested above, some of these less-informative items could be considered for further revision or even deletion (provided it does not uniquely contribute to an important aspect of the definition of the construct).

Item 4, "peer rejection," seems to be particularly problematic in a number of ways. First, researchers have established that this item loads on both the externalizing and internalizing constructs and is not easily modeled within a CFA factor structure (Lane, Oakes et al., 2012;

Wilcox, 2016). This demonstrates this item's decreased ability to provide unique information relating to each specific construct. Additionally, while the "stealing" item loads well on the externalizing factor using CFA (Lane, Oakes, et al., 2012; Wilcox, 2016), the item overall and, specifically, the highest category response option (i.e., "frequently") was rarely endorsed (n = 8). The results of these analyses using the NRM demonstrate an empirical way to collapse such category response options and make the item more informative and better able to discriminate among individuals than it would have been otherwise.

The IRT analyses show further evidence that the peer rejection item provides little measurement information in relation to the other items in the scale. Even with improved item category functioning (i.e., as a dichotomous item) this item provides little measurement precision and psychometric information. This may be due to the nature of this item compared with the other eleven items on the scale. "Peer rejection" could be ambiguous in its interpretation. It could mean either a student's rejection of other peers or it could be interpreted as peers' rejection of that student. Additionally, peer rejection is the only item in which the teacher is asked to judge a student's relationship with others, which may be challenging for middle school teachers who may not have as much opportunity to observe students' peer relationships as elementary school teachers. All other items on the scale deal with the student's individual behavior without consideration for peer or adult relationships. Therefore, careful consideration of the externalizing and internalizing constructs and whether to revise or delete this item seems appropriate in this middle school context.

Practical Significance

Preston and Reise (2015) note the importance of determining practical as well as statistical significance in examining research as it relates to using the NRM in evaluating item

and category functioning on a scale like the SRSS-IE. They note the important impact of varying CBD parameters on the psychometric information provided by items and category response options. They also indicate the importance of having category response options that function as ordered categories because in scales like the SRSS-IE, scoring relies on "consecutive integer weighting" and unordered category response options could result in over- or under-classifying students as at-risk for EBD (Preston & Reise, 2015, p. 403).

However, these researchers also raise the question of whether an increase in model fit (i.e., improved model fit with the NRM as opposed to the GPCM) merits the sacrifice in parsimony. In this study, post-hoc correlations of total scores were done for the externalizing and internalizing subscales for the original item configurations with the revised item and category response configurations suggested here. The correlations between the original item total scores and revised item total scores were very strong for both subscales; therefore, the impact and the practicality of either revising the SRSS-IE items or recoding category responses during the process of scoring the SRSS-IE is in question. Preston and Reise (2015), however, suggest that applying the NRM in examining the psychometric properties of a scale is still beneficial for several reasons: (a) it provides useful psychometric information regardless of whether it is the final model used in the data analyses, (b) the importance of identifying CBD variation "depends on the size and configuration of that variation," and (c) choice of IRT model is ultimately a judgment call made by the researcher (p. 404). Multiple IRT models should be fit to the data and item parameters and graphical depictions of items functioning should be carefully compared (Preston & Reise, 2015). Additionally, comparison of the variance of the total scores on both the original and revised scales here reveals a reduction in the "nuisance variation" described by Preston and Reise (2015) caused by non-informative CBD parameters.

Limitations

Limitations of this research include the difficulty middle school teachers may have rating students with whom they have limited direct contact, especially compared to elementary school teachers. Teacher ratings of students may be based on situational behaviors or they may not be able to view the entire range of a students' behavior in just the middle school homeroom environment.

Another limitation of this study involves the difficulty of measuring psychological constructs in general. Possible difficulties with items on the SRSS-IE may be related the item stems as well as the item options. Using the NRM approach to this research cannot take into account problems with the item stems or operational definitions of these psychological constructs. Lastly, there are also often difficulties in using IRT to measure psychological constructs because such constructs are often representative of atypical or maladaptive behaviors and therefore not normally distributed in the population (Preston & Reise, 2014). Distributions of some psychological constructs, including EBD risk, tend to be skewed toward the extreme ends of the rating scales. In order for IRT to produce non-biased parameter estimates, the data must be normally distributed.

The distribution of data in this research is slightly skewed for some of the items on the SRSS-IE. Some research suggests that the use of the NRM with non-normal data will bias the CBD parameter estimates, although sample size is a major factor in the degree of parameter estimation bias (Preston & Reise, 2014; see also de Ayala & Sava-Bolesta, 1999). For example, Preston & Reise (2014) conducted a simulation study of the effect of sample size of non-normally distributed data on CBD parameter estimates. The bias introduced by non-normally distributed data was much more pronounced for smaller samples sizes (i.e., N=500) than for a

large sample size (i.e., N=2,000; Preston & Reise, 2014)). Given the large sample size in this research study, the potential effects of CBD parameter estimation bias is decreased; however, it remains unclear what the effect would be on CBD parameter estimates for the items on the SRSS-IE in this study.

Recommendations for Practice and Research

This research study can inform both practice and research related to the SRSS-IE. Practitioners (e.g., teachers, school psychologists, counselors) should recognize the importance of ensuring properly ordered category response options when scoring the SRSS-IE and accounting for the natural weighting of each response option. This can be done be revising the SRSS-IE items and category response options to reflect the suggestions made here. It could also be done by re-coding item responses after data is gathered. This would help practitioners to make better-informed decisions regarding which students may need further help or interventions aimed at greatest risk for EBDs. Additionally, practitioners may need further training or clarification on how to interpret specific items on the SRSS-IE (e.g., "peer rejection").

Future research should include having multiple ratings for each student from middle school teachers in various class periods throughout the day to ensure that ratings represent the full spectrum of students' behavior. Such research should also introduce questions about nesting effects and could be addressed using multilevel IRT. Ramsay Curve IRT (RC-IRT) methods could be used to address non-normal distributions of the data using newly developed software such as EQSIRT (Multivariate Software, 2010) that can use RC-IRT with the NRM to estimate parameters (Preston & Reise, 2014).

Future research with the recommendations to changes in item option configurations made in this study should include redesigning the instrument and retesting the scale with the newly

configured items. Items should be reevaluated according to the procedures done in this research to ensure that the recommendations here provide the additional measurement precision and information of each item as described in this study. Other item option configurations might also be considered in revising this scale (i.e., using three response options for all items or using two response options for all items) considering it is atypical to have items on a scale with different numbers of response categories. Lastly, researchers should develop alternative scoring methods and appropriate cut scores based on this new configuration of items and item options.

Conclusion

In conclusion, the use of IRT can be informative in establishing the psychometric soundness of the SRSS-IE when used in middle schools. Particularly, the NRM used in this research to examine the psychometric properties of the SRSS-IE can provide useful information regarding how individual item category options are functioning and the degree to which these categories can meaningfully distinguish among frequencies of behaviors. Such meaningful distinctions are crucial in this context of screening students who may be at risk for EBD. The better these instruments are at detecting problematic behavior in middle school students, the more likely these students are to get the help that they need to be successful in school.

References

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA School-Age Forms & Profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families.

Akaike, H. (1974). A new look at the statistical identification model. *IEEE Transaction Automatic Control, 19, 716-723.*

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

Bartholomew, D., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach* (3rd ed.). London, UK: Wiley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Bock, R. D. (1997). The nominal categories model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 33-49). New York, NY: Springer.

Cai, L. (2013). flexMIRT® version 2: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

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*, 105-117.

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

de Ayala, R., J., & Sava-Bolesta, M. (1999). Item parameter recovery for the nominal response model. *Applied Psychological Measurement, 23*, 3-19.

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

Forness, S. R., Freeman, S. F., Paparella, T., Kauffman, J. M., & Walker (2012). Special education implications of point and cumulative prevalence for children with emotional and behavioral disorders. *Journal of Emotional and Behavioral Disorders, 20*, 4-18.

Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology, 45*, 117-135.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire (SDQ). *Journal of the American Academy of Child and Adolescent Psychiatry, 40*, 1337-1345.

Hardy, C. L., Bukowski, W. M., & Sippola, L. K. (2002). Stability and change in peer relationships during the transition to middle-level school. *The Journal of Early Adolescence, 22*, 117-142.

Hayden, E. P., & Mash, E. J. (2014). Child psychopathology: A developmental-systems perspective. In E. J. Mash & R. A. Barkley (Eds.), *Child psychopathology* (3rd ed., pp. 3-72). New York, NY: The Guilford Press.

Kalberg, J. R., Lane, K. L., Driscoll, S., & Wehby, J. (2011). Systematic screening for emotional and behavioral disorders at the high school level: A formidable and necessary task. *Remedial and Special Education, 32*, 506-520.

Kauffman, J. M., & Landrum, T. J. (2009). *Characteristics of emotional and behavioral disorders of children and youth* (9th ed.). Columbus, OH: Merrill.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of *DSM-IV* disorders in the national comorbidity survey replication. *Arch Gen Psychiatry, 62*, 593-603.

Krueger, R. F., & Markon, K. E. (2006). Reinterpreting comorbidity: A model-based approach to understanding and classifying psychopathology. *Annual Review of Clinical Psychology, 2*, 111-133.

Lane, K. L., Bruhn, A. L., Eisner, S. L., & Kalberg, J. R. (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*, 211-224.

Lane, K. L., Kalberg, J. R., Lambert, W., Crnobori, M., & Bruhn, A. (2010). A comparison of systematic screening tools for emotional and behavioral disorders: A replication. *Journal of Emotional and Behavioral Disorders, 18*, 100-112.

Lane, K. L., Little, M. A., Casey, A. M., Lambert, W., Wehby, J. H., Weisenbach, J. L., & Phillips, A. (2009). A comparison of systematic screening tools for emotional and behavioral disorders: How do they compare? *Journal of Emotional and Behavior Disorders, 17*, 93-105.

Lane, K. L., Menzies, H. M., Oakes, W. P., & Kalberg, J. R. (2012). *Systematic screenings of behavior to support instruction: From preschool to high school*. New York, NY: The Guilford Press.

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*, 244-270.

Lane, K. L., Oakes, W. P., Cantwell, E. D., Menzies, H. M., Schatschneider, C., Lambert, W., & Common, E. A. (2016). Psychometric evidence of the SRSS-IE scores in middle and high schools. *Journal of Emotional and Behavioral Disorders,* 1-13. doi:10.1177/1063426616670862

Lane, K. L., Oakes, W. P., Cantwell, E. D., Schatschneider, C., Menzies, H., Crittenden, M., & Messenger, M. (2016). Student Risk Screening Scale for Internalizing and Externalizing Behaviors: Preliminary cut scores to support data-informed decision making in middle and high schools. *Behavioral Disorders, 42*, 271-284.

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 middle school level. *Assessment for Effective Intervention, 39*, 24-38.

Lane, K. L, Oakes, W. P., Common, E. A., Zorigian, K., Brunsting, N. C., Schatschneider, C. (2015). A comparison between SRSS-IE and SSiS-PSG scores: Examining convergent validity. *Assessment for Effective Intervention, 40*, 114-126.

Lane, K. L., Oakes, W. P., Ennis, R. P., Cox, M. L., Schatschneider, C., & Lambert, W. (2011). Additional evidence for the reliability and validity of the Student Risk Screening Scale at the high school level: A replication and extension. *Journal of Emotional and Behavioral Disorders, 21*, 97-115.

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*, 99-122.

Lane, K. L., Oakes, W. P., Lusk, M. E., Cantwell, E. D., Schatschneider, C. (2015). Screening for intensive intervention needs in secondary schools: Directions for the future. *Journal of Emotional and Behavioral Disorders,* 1-14. doi:10.1177/1063426615618624.

Lane, K. L., Oakes, W. P., Swogger, E. D., Schatschneider, C., Menzies, H. M., & Sanchez, J. (2015). Student Risk Screening Scale for Internalizing and Externalizing Behaviors: Preliminary cut scores to support data-informed decision making. *Behavioral Disorders, 40*, 159-170.

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*, 209-222.

Lane, K. L., Wehby, J., & Barton-Arwood, S. M. (2005). Students with and at risk for emotional and behavioral disorders: Meeting their social and academic needs. *Preventing Social Failure, 49*, 6-9. http://dx.doi.org/10.3200/PSFL.49.2.6-9

Lilienfeld, S. O. (2003). Comorbidity between and within childhood externalizing and internalizing disorders: Reflections and directions. *Journal of Abnormal Child Psychology, 31*, 285-291.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*, 85-106.

McDermott, P. A. (1993). National standardization of uniform multisituational measures of child and adolescent behavior pathology. *Psychological Assessment, 5*, 413-424.

McDermott, P. A., & Weiss, R. V. (1995). A normative typology of healthy, subclinical, and clinical behavior styles among American children and adolescents. *Psychological Assessment, 5*, 162-170.

Multivariate Software (2010). EQSIRT: Item response theory software. Encino, CA: Author.

Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153-164). New York, NY: Springer.

Murray, A. L., Booth, T., & Molenaar, D. (2016). When middle really means "top" or "bottom": An analysis of the 16PF5 using Bock's nominal response model. *Journal of Personality Assessment, 98*, 319-331.

Preston, K. S. J. (2014a, April). *Advanced topics in IRT: Evaluating the effectiveness of each response option with the nominal response model*. PowerPoint presentation at the 94th annual convention of the Western Psychological Association, Portland, OR.

Preston, K. S. J. (2014b, April). *Advanced topics in IRT: Evaluating the effectiveness of each response option with the nominal response model*. PowerPoint presentation at the 94th annual convention of the Western Psychological Association, Portland, OR. Retrieved from http://hssfaculty.fullerton.edu/psychology/kpreston/Plotting.txt

Preston, K. S. J. (2014c, April). *Advanced topics in IRT: Evaluating the effectiveness of each response option with the nominal response model*. PowerPoint presentation at the 94th annual convention of the Western Psychological Association, Portland, OR. Retrieved from http://hssfaculty.fullerton.edu/psychology/kpreston/Wald.txt

Preston, K. S. J., & Reise, S. P. (2014). Estimating the nominal response model under nonnormal conditions. *Educational and Psychological Measurement, 74*, 377-399.

Preston, K. S. J., & Reise, S. P. (2015). Detecting faulty within-item category functioning with the nominal response model. In S. P. Reise & D. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 386-405). New York, NY: Routledge.

Preston, K. S. J., Reise, S. P., Cai, L., & Hays, R. D. (2011). Using the nominal response model to evaluate response category discrimination in the PROMIS Emotional Distress item pools. *Educational and Psychological Measurement, 71*, 523-550.

R Core Team. (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R- project.org/.

Reddy, L. A., Newman, E., De Thomas, C. A., & Chun, V. (2009). Effectiveness of school-based prevention and intervention programs for children and adolescents with emotional disturbance: A meta-analysis. *Journal of School Psychology, 47*, 77-99.

RStudio Team (2015). *RStudio: Integrated development for R*. RStudio, Inc., Boston, MA. Retrieved from http://www.rstudio.com/.

Samejima, F. (1996). Evaluation of mathematical models for ordered polychotomous responses. *Behaviormetrika, 23*, 17-35.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464.

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

Siedman, E., Allen, L., Aber, J. L., Mitchell, C., & Feinman, J. (1994). The impact of school transitions in early adolescence on the self-system and perceived social context of poor urban youth. *Child Development, 65*, 507-522.

Wagner, M., Friend, M., Bursuck, W. D., Kutash, K., Duchnowski, A. J., Sumi, W. C., & Epstein, M. H. (2006). Educating students with emotional disturbances: A national perspective on school programs and services. *Journal of Emotional and Behavioral Disorders, 14*, 12-30.

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

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annuals of Mathematical Statistics, 16*, 117-186. doi:10.1214/aoms/1177731118

Walker, H. M., & Severson, H. H. (1992). *Systematic Screening for Behavior Disorders: Technical manual.* Longmont, CO: Sopris West.

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

Wilcox, M. P. (2016). *Evidence for the validity of the Student Risk Screening Scale in middle school: A multilevel analysis* (Unpublished doctoral dissertation). Brigham Young University, Provo, UT.

Young, E. L., Caldarella, P., Richardson, M. J., & Young, K. R. (2012). *Positive Behavior Support in Secondary Schools: A Practical Guide*. New York, NY: The Guilford Press.

ABSTRACT – Study 2

This research study examined gender differences in the psychometric properties of the *Student Risk Screening Scale for Internalizing and Externalizing Behaviors* (SRSS-IE; Lane, Oakes, et al., 2012) using Item Response Theory (IRT) methods among a sample of 2,122 middle school students. The SRSS-IE is a recently revised screening instrument aimed at identifying students who are potentially at risk for emotional and behavioral disorders (EBD). Differential item functioning (DIF) and differential step functioning (DSF) methods were used to examine differences in item and response option functioning according to student gender variables. Additionally, test information functions (TIFs) were used to determine whether preliminary recommendations for moderate EBD risk and high EBD risk cut scores differ by gender. Results of this study indicated that two of the items on the SRSS-IE systematically favor males over females and one item systematically favors females over males. Additionally, examination of TIFs demonstrated different degrees of measurement precision at various levels of theta for males and females on both the externalizing and internalizing constructs. Implications of these results are discussed in relation to possible revisions of the SRSS-IE items, cut scores, or scale scoring procedures.

Keywords: Student Risk Screening Scale, emotional and behavioral disorders, universal screening, differential item functioning, cut scores

Gender Differences in the Student Risk Screening Scale for Internalizing and Externalizing Behaviors: An Examination of Differential Item Functioning and Cut Scores

Purpose of Study

This study is an extension of a recent examination of the psychometric properties of the *Student Risk Screening Scale for Internalizing and Externalizing Behaviors* (SRSS-IE) among a sample of middle school students. The research reported here specifically examined the degree of gender Differential Item Functioning (DIF) and Differential Step Functioning (DSF) of the items on the SRSS-IE. Additionally, the pre-established cut scores for the SRSS-IE were examined to determine whether these cut scores are appropriate based on gender. The *Student Risk Screening Scale* (SRSS; Drummond, 1994) is the precursor to the SRSS-IE and was developed to screen for patterns of antisocial behavior (Lane, Menzies, Oakes, & Kalberg, 2012).

Originally, the SRSS focused on identifying students' behavior patterns manifest as externalizing behaviors. Externalizing behaviors can be conceptualized as misbehavior directed toward others (e.g., cheating, low academic achievement, aggressive behaviors) and such behaviors are often associated with an increased risk for Emotional and Behavioral Disorders (EBD; Lane, Menzies, Oakes, & Kalberg, 2012).

Lane, Oakes, and colleagues (2012) extended the constructs related to EBD risk identified in the SRSS by developing additional items to address internalizing behavior patterns (e.g., feeling anxious, shy, withdrawn). In doing so, they hoped to identify students who could be at risk for a broader range of EBD (Lane, Oakes, et al., 2012). The purpose of this study was to examine whether both the externalizing and internalizing items on the SRSS-IE perform as expected by examining DIF, DSF, and using Item Response Theory (IRT) to explore whether

pre-established cut scores provide the same amount of measurement information at a given level for both males and females.

Some research has been done to broadly examine the psychometric properties of the SRSS-IE (e.g., Lane, Oakes, et al., 2012) and recent research by Lane, Oakes, Cantwell, Schatschneider, and others (2016) has focused on examining cut scores at the middle school level. However, a DIF analyses of these items has not been conducted at either the elementary or middle school levels. A careful examination of potential gender DIF and DSF effects is important in order to ensure that items are functioning consistently across gender groups. Additionally, the use of IRT has not been applied as a methodology to examine current cut scores and to determine whether these cut scores can best identify those at high risk of externalizing and internalizing behavioral disorders in both males and females.

This research study aims to provide evidence of the measurement precision the SRSS-IE in its ability to accurately identify externalizing and internalizing behaviors of both male and female students. The purpose in attempting to identify students who may be at risk for EBD is to eventually help them to receive additional, targeted support in schools. Additionally, DIF and DSF analyses of the SRSS-IE data in this context could provide evidence of validity that the instrument is working as intended among both males and females during early adolescence.

Review of Literature

Universal Screening in Schools

Part of a successful Tier 1 (universal or school-wide) intervention is an accurate measurement program from which information about students can be obtained and used for meaningful purposes, such as improved classroom instruction (Lane, Menzies, Oakes, & Kalberg, 2012). Screening at the Tier 1 level is one way that teachers and administrators can

obtain such information and which can then be used to take a preventative, rather than punitive, approach to behavioral concerns. In order to draw meaningful conclusions from screening instruments, however, it is essential that these instruments are psychometrically sound (Glover & Albers, 2007). Lane, Oakes, Menzies, and Kalberg (2012) emphasize that psychometric soundness is essential if resulting data from such measurements are used to make important decisions about students. They reiterate that screening instruments should be reliable and valid and that measurement systems should be implemented in schools with fidelity, thus ensuring that these tools are used carefully to help address the appropriate level of risk associated with students' psychological needs (Lane, Oakes, Menzies, & Kalberg, 2012).

An important area of focus for screening is to identify students who may be at risk for emotional and behavioral disorders (EBD). EBD risk can be manifest as specific actions related to externalizing behaviors (e.g., behavior problems, delinquency) or as internalizing behaviors (e.g., social isolation, anxiety; Lane, Menzies, Oakes, Kalberg, 2012; Lane, Oakes, Lusk, Cantwell, & Schatschneider, 2016). The relationship between externalizing and internalizing behaviors in the context of EBD risk has been extensively examined in the research literature (e.g., Achenbach & Rescorla, 2001; Forness, Freeman, Paparella, Kauffman, & Walker, 2012; Krueger & Markon, 2006; Hayden & Mash, 2014).

The externalizing construct has been operationalized and categorized by researchers to include specific, observable behaviors (e.g., Achenbach & Rescorla, 2001). These behavior descriptions were used as the basis for the SRSS and subsequently for the externalizing portion of the SRSS-IE. The seven externalizing behaviors or problems included on the SRSS-IE include the following: (a) stealing; (b) lying, cheating, sneaking; (c) behavior problems; (d) peer rejection; (e) low academic achievement; (f) negative attitude; and (g) aggressive behaviors

(Drummond, 1994; Lane, Oakes et al., 2012). Externalizing behaviors can also be conceptualized as behaviors directed toward others or undercontrolled problems (Hayden & Mash, 2014).

While the internalizing construct has been more difficult to operationalize in the research literature since behaviors associated with this construct are less observable, these behaviors are sometimes viewed as inner-directed behaviors or over controlled problems (Hayden & Mash, 2014; Lane, Menzies, Oakes, & Kalberg, 2012). Despite this difficulty in operationalizing internalizing behavior problems, Lane, Oakes, and colleagues (2012) concluded that some behaviors or problems that should be included on the SRSS-IE in the internalizing dimension are the following: (a) emotionally flat; (b) shy, withdrawn; (c) sad, depressed; (d) anxious; and (e) lonely. The addition of internalizing behaviors with externalizing behaviors in screening students who may be at risk of EBD is an essential component to better capturing the theoretical construct of EBD risk.

While externalizing behaviors and internalizing behaviors may be conceptualized as separate constructs, they are also related to one another in important ways (McDermott, 1993; McDermott & Weiss, 1995). Some researchers report a relatively high level of comorbidity between these two constructs with correlations as high as $r = .5$ (Krueger & Markon, 2006). In relation to the SRSS-IE, this demonstrates the importance of inclusion of both the externalizing behavior and internalizing behavior constructs into accurate measurement of EBD risk (Lane, Oakes, et al., 2012).

These two constructs (i.e., externalizing behaviors and internalizing behaviors) are of interest in this research study since they are both important components in carefully measuring EBD risk. It is crucial to screen students for EBD risk in middle schools since students of

middle school age are particularly at risk for some form of EBD with estimates as high as 20% of youth who are at risk (Forness, et al., 2012). EBD can impair student learning, damage relationships with peers, and hinder students' abilities to successfully complete high school graduation requirements (Lane, Menzies, Oakes, & Kalberg, 2012; Lane, Oakes, Lusk, et al., 2016; Wagner et al., 2006). Establishing the psychometric soundness of the SRSS-IE to precisely identify students most likely to be at risk for EBD in middle schools could help alleviate such academic and social difficulties and enable school teachers and leaders to provide early interventions best suited to individual students' needs.

Screening Instruments

Various instruments have been developed as screening tools to attempt to identify students at risk for EBD. Some of these instruments include the Systematic Screening for Behavior Disorders (SSBD; Walker & Severson, 1992; see also Walker, Severson, & Feil, 2014), the BASC-2 Behavioral and Emotional Screening System (BASC-2 BESS; Kamphaus & Reynolds, 2007), and the Student Risk Screening Scale (SRSS; Drummond, 1994). A systematic comparison of some of these screening tools in correctly identifying students at risk for EBD was done by Lane and colleagues (2009). These researchers also examined these instruments in terms of ease of administration. Overall, the psychometric properties of these instruments were strong and closely matched the then gold standard instrument, the SSBD (Kauffman & Landrum, 2009; Lane et al., 2009). However, the researchers in this study also noted that these screening tools were somewhat lacking in terms of their measurement precision in identifying at risk students who exhibited internalizing behavior patterns (Lane et al., 2009). These findings were primarily based on screening done at the elementary level. The comparability of some of these

screening tools for middle and high school students is still emerging (Caldarella, Young, Richardson, Young, & Young, 2008; Kalberg, Lane, Driscoll, & Wehby, 2011).

A replication of the Lane and colleagues' (2009) study was done to compare the SSBD and the SRSS and yielded similar results (Lane, Kalberg, Lambert, Crnobori, & Bruhn, 2010). These results included evidence of internal consistency reliability and a high degree of sensitivity and specificity in identifying students who were at risk for EBD, at least for those students who were perceived as exhibiting externalizing behaviors. For those students who were perceived as exhibiting internalizing behaviors, however, the SRSS was inadequate at fully capturing those students' behavior challenges (Lane, Kalberg, Lambert, Crnobori, & Bruhn, 2010).

Lane and colleagues' (2010) study demonstrated the importance of including the theoretical construct of internalizing behaviors in the identification, and subsequent treatment, of EBD. The SRSS-IE was therefore created as an extension of the original SRSS developed by Drummond (1994) and included items aimed at identifying internalizing behaviors of students at school (Lane, Oakes et al., 2012). The structure of the SRSS-IE includes indicators of EBD risk that fall distinctly into the two separate theoretical constructs of externalizing and internalizing behaviors. This theoretical structure is intended to capture a broader range of indicators that better represents both the breadth and depth of factors relating to EBD risk.

Changing the items of an instrument to match an underlying theoretical structure necessitates a re-examination of the psychometric soundness of the instrument. The SRSS-IE has been researched in several studies to evaluate its psychometric properties, including its latent two-factor structure (Lane, Menzies, Oakes, Lambert et al., 2012; Lane, Oakes et al., 2012; Lane, Oakes, Carter, Lambert, & Jenkins, 2013; Lane, Oakes, Lusk, et al., 2016). However, more research needs to be done to explore whether each item on the SRSS-IE is functioning as

intended for specific demographic groups (e.g., gender, ethnicity) and whether the two theorized constructs are functioning as distinct constructs for those same demographic groups.

Additionally, the median age for onset of various anxiety and impulse-control disorders has been indicated in research to be approximately age 11 with the majority of emotional and behavioral difficulties manifest by age 14; therefore, evidence of validity of the SRSS-IE is critical among middle school and high school students (Kessler et al., 2005). The psychometric properties of SRSS have been carefully researched at the elementary, middle school, and high school levels (see Lane et al., 2010; Lane et al., 2011; Lane, Bruhn, Eisner, & Kalberg, 2010). The SRSS-IE, on the other hand, has been primarily researched at the elementary level (see Lane, Menzies, Oakes, Lambert et al., 2012; Lane, Oakes et al., 2012; Lane, Oakes, Lusk, et al., 2016) with some recent exceptions that have examined the SRSS-IE at the middle and high school levels (e.g., Lane et al., 2013; Lane, Oakes, Lusk, Cantwell, & Schatschneider, 2016). Additional research examining the psychometric functioning of the SRSS-IE in middle level grades is a critical component to ensuring that the SRSS-IE can appropriately and accurately identify students who may be at risk for EBD.

Differential Item Functioning

Detecting the existence and the degree of DIF is important in this study because of previous research highlighting gender differences related to mental health concerns. Historically, males have been more likely to show outwardly aggressive behaviors and girls have been more likely to have internalizing disorders. For example, boys are at least twice as likely to have disorders demonstrating externalizing symptoms such as ADHD (Nigg & Barkley, 2014). Females are more likely to exhibit higher rates of internalizing symptoms than boys such

as in anxiety disorders (Higa-McMillan, Francis, & Chorpita, 2014) and depressive disorders (Hayden & Mash, 2014).

Some researchers have found that although there have been higher rates of boys with externalizing problems and girls with internalizing problems in children who were referred for treatment, gender differences in this spectrum were actually small for children who were not referred for treatment (Achenbach, Howell, Quay, Conners, & Bates, 1991). However, other research suggests that gender differences do exist in terms of risk factors, prevalence rates, and how some mental disorders develop or change over time (Young, Sabbah, Young, Reiser, & Richardson, 2010). The latter research study reported that males were almost three times as likely to be nominated by teachers in the context of internalizing and externalizing behaviors. Any discrepancies in reports of gender differences could be due to actual differences between the genders or to differences in the way the constructs are measured. A gender DIF analysis of the items on the instrument could provide empirical evidence to enhance understanding of which is more plausible in the data in this study related to the SRSS-IE.

In addition to determining which items may exhibit a greater degree of gender DIF, DSF can also be used with polytomous items to specifically determine at which option level the gender differences are occurring (Penfield & Gattamorta, 2009). This is particularly important in scales using polytomous items because of the impact of each item's contribution to a student's score, and, in this case, on the degree of EBD risk according to the SRSS-IE (Penfield & Gattamorta, 2009). Any identified differential functioning according to students' gender could provide a source of potential bias in the instrument's abilities to correctly identify, and subsequently help, students at risk for EBD.

Determining Cut Scores using IRT

Gender differences may also play a part in whether pre-established cut scores can correctly separate both males and females who may be at risk for EBD. Setting appropriate cut scores for the SRSS-IE is of particular interest in the current research study. Research on the SRSS-IE has employed the use of receiver operating characteristic (ROC) curves to determine cut scores (e.g., Lane, Oakes, Swogger, et al., 2015). This approach to setting cut scores "can be conceptualized as an array of possible cutting scores, with each cut score offering a unique balance of benefit (sensitivity) and cost (specificity)" (Lane, Oakes, Swogger, et al., 2015, p. 166). Setting cut scores for the SRSS-IE using this method was done by deciding what that balance should be given the cost of over- or under-identifying students with problem behaviors (Lane, et al., 2009).

An empirical approach to setting cut scores can also be done using IRT methods (de Ayala, 2009). An IRT analysis of setting cut scores has not been done with the SRSS-IE. While some research suggests that conclusions on appropriate cut scores from ROC curve analysis closely mirrors conclusions drawn from IRT analyses (DiStefano & Morgan, 2011; Yovanoff & Squires, 2006), it is important to test this hypothesis with the SRSS-IE and specific age groups including those with differing demographic variables. Additionally, using IRT to establish cut scores on a screening instrument has several advantages including allowing persons and scores to be on a single, equal-interval scale (i.e., a logit scale). IRT modelling techniques are also sample independent and do not require a previously diagnosed group such as in a ROC curve analysis (DiStefano & Morgan, 2011).

Given potential differences in ratings between the genders as described above, it is important to determine whether the cut scores should be the same for both genders.

Additionally, IRT analyses should be employed for examining the internalizing construct in particular since research to establish a cut score for this construct on the SRSS-IE is limited (Lane, Oakes, Lusk, et al., 2016; Lane, Oakes, Swogger, et al., 2015). Establishing appropriate cut scores is essential to removing potential bias that could result in incorrectly identifying students as either at-risk for EBD when they are not, or, more seriously, failing to identify students as at-risk for EBD when they are (Severson, Walker, Hope-Doolittle, Kratochwill, & Gresham, 2007). Lastly, it should be examined in this research whether the cut scores established by Lane, Oakes, and colleagues (2012) in an elementary school context are also appropriate for middle school students in this research study. While research is ongoing in establishing such cut scores for a middle school population (Lane, Oakes, Cantwell, Schatschneider, et al., 2016), this research can contribute to the body of literature on the subject by providing additional evidence for determining appropriate cut scores and to consider if the evidence for the cut scores is consistent for all gender groups.

In summary, the SRSS-IE has shown promise as an important tool for screening school children and adolescents who may be at risk for EBD. Further research, however, is critical in providing evidence of the psychometric soundness of this instrument. Such research should include the use of current psychometric methodologies, such as IRT, to get a more accurate representation of how individual items are performing and if they are performing as intended, especially among both males and females. Ultimately, the goal of this research in attempting to provide validity evidence for the SRSS-IE is to make the instrument psychometrically sound and more widely available in the field for the benefit of teachers and students.

Method

Participants and Setting

This study focused on two research questions:

1. Which, if any, of the items or item response options function differentially for males and females?
2. What evidence is there to support using the cut scores designated in previous research? Should those cut scores be the same or different for males and females?

In order to help answer these research questions, de-identified, archival data from 2,122 students from three middle schools were analyzed. The data were originally generated by 93 teachers (61% female) who completed the SRSS-IE for each student in their first period (homeroom) class.

The SRSS-IE data analyzed in this study were collected previously from the teachers and the psychometric analyses were done on the existing data set. Demographic information about the students' gender and race, and the teachers' demographic characteristics were collected along with the SRSS-IE data (Wilcox, 2016). The data were initially gathered from a school district in the mountain west. The data came from three middle schools within the district and represent information about students in grades 6 through 8. All available data were used for this study. Grade and gender information for students is displayed in Table 1. Table 2 contains demographic information about racial backgrounds for both teachers and students in this research study.

Table 1

Frequency Distribution of Students' Gender and Grade

	Male		Female		
Grade	n	%	n	%	Total
6	372	47%	420	53%	792
7	364	54%	311	46%	675
8	346	53%	309	47%	655
Total	1082	51%	1040	49%	2122

Table 2

Frequency Distribution of Teachers' and Students' Racial Backgrounds

	Teachers (n = 93)		Students (n = 2,122)	
Race	n	%	n	%
White	64	69%	1664	78%
Hispanic/Latino	16	17%	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%

Measure

The SRSS-IE consists of 12 items including seven items representing the externalizing construct and five items representing the internalizing construct in which a teacher rates each student on the frequency of each observed behavior. The frequency scale consists of four response options ranging from 0 (*never*), 1 (*occasionally*), 2 (*sometimes)*, to 3 (*frequently*). The seven items representing externalizing behaviors include the following: (a) stealing; (b) lying, cheating, sneaking; (c) behavior problems; (d) peer rejection; (e) low academic achievement; (f) negative attitude; and (g) aggressive behaviors. The five items representing internalizing behaviors include the following: (a) emotionally flat; (b) shy, withdrawn; (c) sad, depressed; (d) anxious; and (e) lonely. The teachers rated each student on each of these 12 behaviors.

Several studies have been conducted to estimate the reliability of SRSS-IE scores and to collect validity evidence. Coefficient alpha reliability estimates for the SRSS-IE items at the elementary level were initially estimated to be .84 for the externalizing behavior items and .72 for the internalizing behavior items (Lane, Oakes et al., 2012). Examination of the factor structure of the data using an exploratory factor analysis (EFA) at the elementary level demonstrated evidence supporting a two-factor model (i.e., externalizing and internalizing) with the peer rejection item loading fairly high on both latent constructs (Lane, Oakes et al., 2012). Evidence of convergent validity of the SRSS-IE with previously existing instruments (e.g., SDQ) has been demonstrated and correlations of the total scores on the SRSS-IE to the SDQ subscales has ranged from .49 to .75 for students at the elementary school level (Lane, Oakes et al., 2012). Subsequent studies of the SRSS-IE have demonstrated similar reliability of scores and validity evidence at the elementary (e.g., see Lane et al., 2010; Lane, Bruhn, Eisner, & Kalberg, 2010), middle school (Lane, et al., 2013), and high school levels (Lane et al., 2011; Lane, Oakes, Cantwell, Menzies, et al., 2016).

Subscale scores for each student on the SRSS-IE are calculated by adding together the teacher's ratings of that student's scores for each of the seven externalizing items to produce an externalizing risk score ranging from 0 to 21. The teacher's ratings for each of the five internalizing items for each student are then added together to yield an internalizing risk score ranging from 0 to 15 (Lane, Oakes, Common, et al., 2015). Students are classified according to the following risk categories for the externalizing subscale: low risk (0-3); moderate risk (4-8); and, high risk (9-21; Lane, Menzies, Oakes, & Kalberg, 2012). While cut scores to determine low, moderate, and high risk for EBD were not initially established for the internalizing subscale (Lane, Oakes, Lusk, et al., 2016), recent studies at the elementary level have proposed that the

total internalizing scores for these cut points should be: low risk (0-1); moderate risk (2-3); and, high risk (4-15; Lane, Oakes, Swogger, et al., 2015). Studies are being conducted regarding possible cut scores for use in classifying at-risk students at the middle school and high school levels (Lane, Oakes, Cantwell, Schatschneider, et al., 2016).

Design and Analyses

To analyze the data for the first research question in this study related to DIF and DSF, *Differential Item Functioning Analysis System* (*DIFAS) 5.0* was used (Penfield, 2013). The whole data set was divided into two data sets, one consisting only of data from the externalizing items on the SRSS-IE, and a separate data set consisting only of data from the internalizing items on the SRSS-IE. The DIF analyses were conducted on each data set separately. Males were used as the focal group and females were used as the reference group since the focal group should be representative of the group being investigated (i.e., in the case of the SRSS-IE, males are typically overrepresented as being at risk for EBD; de Ayala, 2009). DIF analyses were conducted to determine which, if any, of the items or item response options function differentially for males and females. Since this instrument theoretically consists of only ordered polytomous items, a DSF analysis was conducted following the traditional DIF analysis (see Penfield, Gattatmorta, & Childs, 2009). The DSF analysis is useful in determining specifically where, within each of the item options, the differential ratings may occur.

The *DIFAS* software produces several statistics to determine the degree of DIF for each item. These statistics are: (a) the Mantel chi-square, (b) the Liu-Agresti cumulative common log-odds ratio (along with its standard error), (c) the standardized Liu-Agresti cumulative common log-odds ratio, (d) Cox's estimator of the multivariate hypergeometric noncentrality

parameter (along with its estimated standard error), and (e) the standardized value of Cox's estimator (Penfield, 2013).

The degree of DIF was evaluated according to the pre-determined criteria for each statistic. For the Mantel chi-square statistic, items with values above 3.84 (indicating a Type I error rate ≤ .05) were considered as evidence of DIF (Mantel, 1963; Zwick, Donoghue, & Grima, 1993; Zwick, Thayer, & Mazzeo, 1997). For the Liu-Agresti cumulative common log-odds ratio (LA-LOR), positive values indicated DIF in favor or the reference group while negative values indicated DIF in favor of the focal group (Liu & Agresti, 1996; Penfield & Algina, 2003). For the standardized Liu-Agresti cumulative common log-odds ratio (LOR Z), items with values greater than 2.0 or less than -2.0 were considered as evidence of DIF (Liu & Agresti, 1996). For Cox's noncentrality parameter estimator (Cox's B), any positive values indicated DIF in favor of the reference group while negative values indicated DIF in favor of the focal group (Camilli & Congdon, 1999). The standardized version of Cox's noncentrality parameter (Cox's Z) provided values for each item where a value greater than 2.0 or less than -2.0 was considered evidence of DIF (Camilli & Congdon, 1999).

The degree of DSF was evaluated using three parameter estimates from the DIFAS output: (a) adjacent categories log-odds ratio (AC-LOR), (b) the standard error estimator of the DSF effect estimate, and (c) the ratio of each DSF effect estimate over its respective standard error estimator (z; Penfield, Gattamorta, & Childs, 2009). Penfield, Gattamorta, and Childs (2009) suggest the following criteria for categorizing the magnitude of DSF effects using the absolute value of the AC-LOR parameter: an AC-LOR < .43 corresponds to a small DSF effect; .43 ≤ AC-LOR < .64 corresponds to a medium DSF effect; and an AC-LOR ≥ .64

corresponds to a large DSF effect. The degree of DSF in this study will be judged based on these criteria.

IRT was used to examine the evidence about the cut scores designated in previous research and whether those cut scores should be the same or different for males and females. The Nominal Response Model (NRM) was selected for this analysis because of its ability to provide item information and category information for each item option within the item (Preston & Reise, 2015). The data analyses were performed using the flexMIRT software (Cai, 2013) as well as R 3.3.1 (R Core Team, 2014) and RStudio 0.99.903 (RStudio Team, 2015). In order to answer the research questions, Test Information Functions (TIFs) were examined to determine at what specific levels of each trait (i.e., externalizing and internalizing) the SRSS-IE provides the most measurement information. Information from the flexMIRT output was used and run through RStudio using code to generate various IRT plots, including TIFs (Preston, 2014).

Separate TIFs were created for males and females together under the externalizing construct, males and females together under the internalizing construct, males and females separately under the externalizing construct, and males and females separately under the internalizing construct. Item option functioning was then examined using the NRM for each item and poorly functioning category options were collapsed into adjacent categories (see Preston, 2014a; Moulton & Young, 2016). The resulting configuration of item options from the previous study is as follows: the items "stealing," "peer rejection," "low academic achievement," "negative attitude," "shy, withdrawn," and "sad, depressed" should be modeled as dichotomous items (i.e., 0 = never does the behavior, 1 = at least sometimes does the behavior); and the items "lying, cheating, sneaking," "behavior problems," "aggressive behaviors," "emotionally flat,"

"anxious," and "lonely" should be modeled as having three item options (i.e., 0 = never does the behavior, 1 = sometimes does the behavior, and 2=frequently does the behavior).

TIFs were then generated again for males and females separately and for both the externalizing and internalizing constructs. The flexMIRT default convergence criteria had to be adjusted (E-step tolerance value was set to $400e^{-2}$) in the configuration file syntax only for the original externalizing items in the female group in order to reach convergence. The data analyses on all other data sets reached convergence without adjusting the default convergence criteria. In each analysis, the peaks of each TIF were examined to see whether they were at or near the cut scores cited in previous research (Lane, Oakes, Lusk, et al., 2016; Lane, Oakes, Swogger, et al., 2015). Where previous cut scores have not been fully established for middle school students, a recommended cut score at the highest level of precision of the instrument (i.e., the peak of the test information function) represented by the corresponding level of endorsability of the items representing each trait was recommended.

Results

Two separate analyses were conducted; the first analysis evaluated the externalizing items and the second analysis considered the internalizing items. The results obtained from the DIFAS output for the DIF analyses are represented in Table 3. Results emphasized in bold demonstrate a relatively high degree of overall item-level DIF. Positive parameter estimates favor the reference group (i.e., females) while negative parameter estimates favor the focal group (i.e., males).

Table 3

Indicators of the Degree of DIF for the 12 Items on the SRSS-IE

			LA-LOR			COX		
#	Item	Mantel	LA-LOR	SE	Z	B	SE	Z
1	Stealing	1.86	-0.34	0.24	-1.43	-0.25	0.18	-1.37
2	Lying, cheating, sneaking	2.90	0.24	0.15	1.66	0.17	0.10	1.70
3	Behavior problems	**31.51**	**-0.85**	**0.15**	**-5.57**	**-0.53**	**0.09**	**-5.61**
4	Peer rejection	0.05	0.03	0.15	0.22	0.02	0.09	0.23
5	Low academic achievem.	**5.39**	**0.28**	**0.12**	**2.37**	**0.16**	**0.07**	**2.32**
6	Negative attitude	0.04	0.03	0.14	0.19	0.02	0.09	0.19
7	Aggressive behaviors	1.75	0.22	0.17	1.28	0.15	0.11	1.32
8	Emotionally flat	1.10	0.15	0.14	1.05	0.10	0.09	1.05
9	Shy, withdrawn	2.23	0.18	0.12	1.50	0.11	0.07	1.50
10	Sad, depressed	**4.58**	**-0.33**	**0.16**	**-2.14**	**-0.24**	**0.11**	**-2.14**
11	Anxious	0.28	-0.08	0.15	-0.53	-0.05	0.09	-0.53
12	Lonely	0.62	-0.12	0.15	-0.79	-0.09	0.12	-0.79

Results of the DSF analyses were also examined with special attention paid to the three items above. The location of the differences within these items was examined using DSF procedures in DIFAS. The results of the analyses for these items are displayed in Table 4.

Table 4

Indicators of the Degree and Location of DSF for Items 3, 5, and 10

Item	Step	AC-LOR	SE	Z	DSF Size
3	**1**	**-0.627**	**0.177**	**-3.535**	**Medium**
3	**2**	**-0.688**	**0.275**	**-2.504**	**Large**
3	3	-0.204	0.398	-0.513	Small
5	1	0.250	0.160	1.565	Small
5	2	0.271	0.204	1.329	Small
5	3	-0.122	0.212	-0.575	Small
10	1	-0.108	0.177	-0.610	Small
10	2	-0.248	0.290	-0.855	Small
10	**3**	**-0.614**	**0.435**	**-1.411**	**Medium**

It should be noted that Item 7 (Aggressive behaviors), as a whole, did not indicate a significant degree of overall DIF; however, the item does display a large DSF effect in favor of the

reference group (i.e., females) between options 2 and 3 (AC-LOR = 0.991, SE = 0.592, Z = 1.674).

For the cut score analyses, the original TIFs using the NRM for the externalizing and internalizing constructs with all males and females included together are represented in Figure 1.

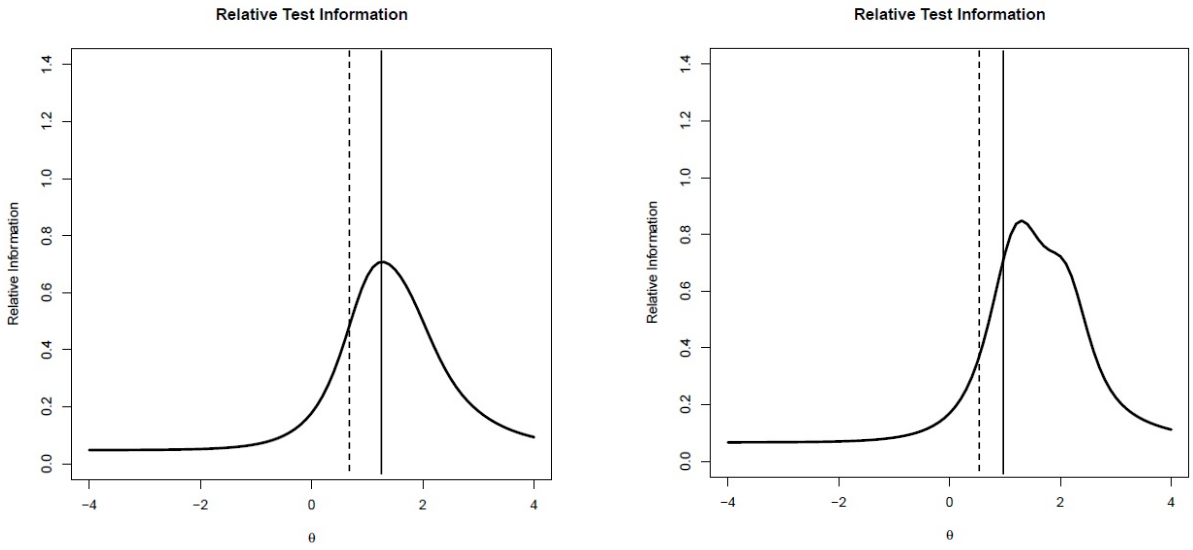

Figure 1. Test Information Functions for the externalizing scale (on the left) and the internalizing scale (on the right). Males and females are included together in both TIFs for comparison purposes. The dotted line represents the cut score for moderate risk while the solid line represents the cut score for high risk of EBD.

The TIF for the externalizing items peaks at a level of theta that appears to be slightly above 1.0. According to current research, a student would be considered at a high risk for EBD on this construct with an overall score of 9 or higher (Lane, Menzies, Oakes, & Kalberg, 2012). According to the flexMIRT output, this score corresponds to a theta of approximately 1.34. The cut score for moderate risk on the externalizing scale is currently set at 4 which corresponds to a theta of approximately .63. The TIF for the internalizing items peaks at a level of theta that

appears to be also slightly above 1.0. According to current research, a student would be considered at a high risk for EBD on this construct with an overall score of 4 or higher (Lane, Oakes, Swogger, et al., 2015). According to the flexMIRT output, this score corresponds to a theta of approximately .90. The cut score for moderate risk on the internalizing scale is currently set at 2 which corresponds to a theta of approximately .37.

The TIFs for both constructs with males and females represented separately and the current recommended cut scores indicated graphically by vertical lines are depicted in Figures 2 and 3. The original scale TIFs are on the left and the revised scale TIFs are on the right.

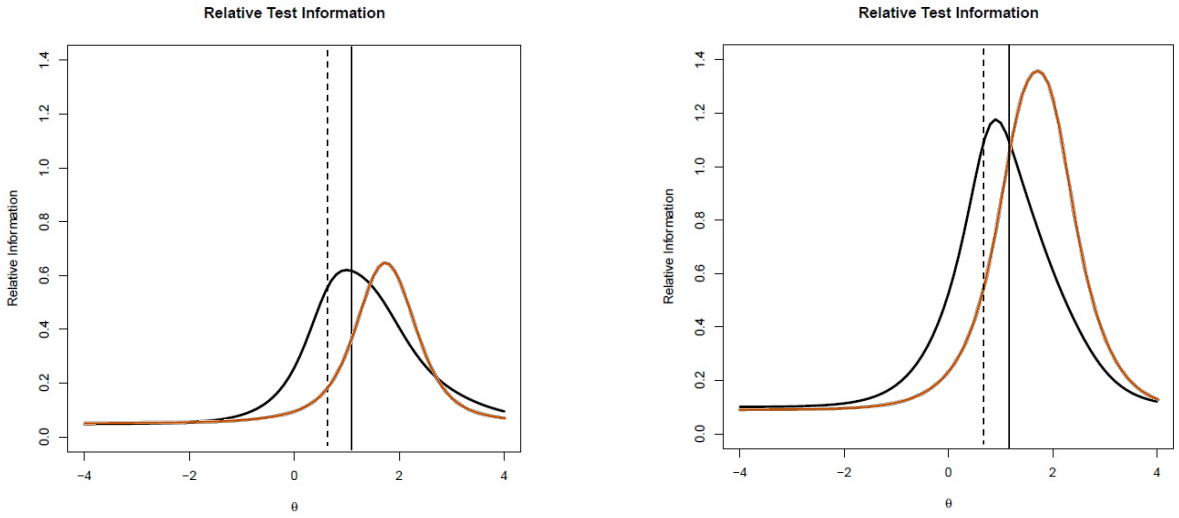

Figure 2. Test Information Functions for the externalizing scale. The TIF for males is represented by the black curve and the TIF for females is represented by the orange curve. The dashed line represents the current cut score for moderate risk at a theta of .63 while the solid line represents the current cut score for high risk at a theta of 1.34 for externalizing behaviors.

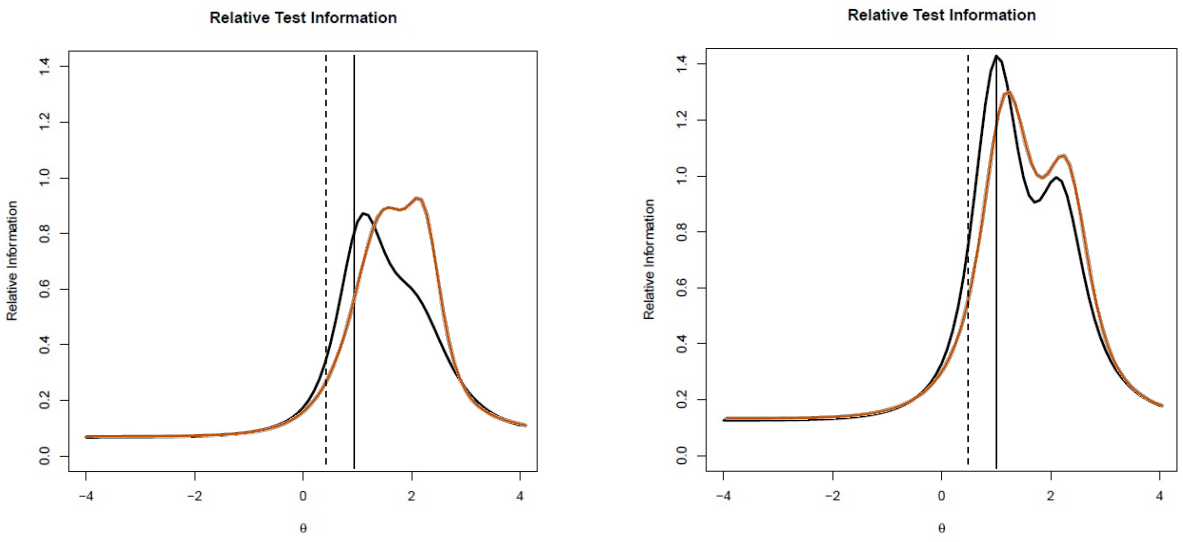

Figure 3. Test Information Functions for the internalizing scale. The TIF for males is represented by the black curve and the TIF for females is represented by the orange curve. The dashed line represents the current cut score for moderate risk at a theta of .37 while the solid line represents the current cut score for high risk at a theta of .90 for internalizing behaviors.

The peak of the TIF for the externalizing items appears to be at a level of theta that is slightly below 1.0 for males and slightly below 2.0 for females. The TIF for the internalizing items peaks at a level of theta that appears to be approximately 2.0 for females and 1.0 for males. If scale scores were used to determine high risk cut scores at the peaks of test information, the cut scores for each scale would be 8 for males and 15 for females for the externalizing subscale. If peak test information was used to determine high risk cut scores for the internalizing subscale, the cut scores would be 4 for males and 10 for females.

Assuming that the cut scores remain at the same levels of theta as originally proposed, the new scale score cutoff for the revised externalizing items would have different cut scores for

males and females and would be at 2 for moderate risk and 4 for high risk for females and 3 for moderate risk and 6 for high risk for males. The new scale score cutoff for the revised externalizing items would have the same cut scores for males and females and would be at 1 for moderate risk and 1 for high risk. However, if the high risk cut scores are instead placed at the peaks of the TIFs where the most measurement information is available, the cut scores for the externalizing items would be 7 for females and 5 for males. For the internalizing items, the peak levels of information represented by the TIFs would place females and males at 3.

Discussion

IRT provides a useful framework to examine gender differences and similarities in the item and test functioning of the SRSS-IE. The results from this study provide insight into how gender may be an important variable to consider when screening is conducted by teachers as well as when the SRSS-IE is scored and decisions are made related to EBD risk.

DIF and DSF on the SRSS-IE

There is evidence to suggest that at least some of the items on the SRSS-IE function differentially for middle-school aged males and females. Particularly, Item 3 (behavior problems), Item 5 (low academic achievement), and Item 10 (sad, depressed) exhibit both a higher degree of overall item DIF and DSF at certain levels. Item 7 (aggressive behaviors) exhibits a large DSF effect at the high end of the trait.

Discovering and measuring degree of DIF and DSF effects is important because these effects demonstrate systematic differences in how the items on the SRSS-IE are answered based on students' gender at a given level of theta, rather than on the level of the latent trait that students manifest. On item 3, for example, teachers systematically favor males over females for the first two steps in frequency of behavior problems (i.e., males are systematically rated higher

on this item at the same level of theta as females). This trend does not extend to the third step, perhaps suggesting that if behavior problems are severe or frequent enough, males and females are rated somewhat equally.

On item 5 (low academic achievement), the DSF is small across all steps. However, overall DIF indicates a favoring of females across the item. This particular item, low academic achievement, should be a relatively objective indicator of student behavior based on student academic performance. The existence of DIF on this item, however, may lead to questions regarding how teachers perceive students' academic abilities which have been shown to differ according to student gender (Pomerantz, Altermatt, & Saxon, 2002). Item 7 (aggressive behaviors) systematically favors females at the highest levels of the trait. Therefore, males are more likely than females to be perceived as displaying frequent aggressive behaviors. Lastly, item 10 (sad, depressed) favors the focal group (i.e., males), particularly at higher levels of the trait, as seen on step 3. This indicates that teachers who rate students on the sad/depressed trait on the SRSS-IE are more likely to rate males as being high relative to females at the same level of the trait. These items that exhibit a larger degree of DIF seem to be items that capture more general behaviors as opposed to discreet, specific behaviors. Additionally, they seem to reflect gender stereotypes (Hoffman, Powlishta, & White, 2004; Young, Sabbah, Young, Reiser, & Richardson, 2010).

Gender Differences in Cut Scores on the SRSS-IE

The research on cut scores suggests that although the current version of the SRSS-IE has a single cut score for males and females in middle schools, the amount of measurement information obtained by this scale differs by gender. If the scale remains unchanged, consideration could be given to having separate cut scores according to gender. The scale does

not have the same amount of measurement precision for males, for example, at the same levels of theta as it does for females on both the externalizing and internalizing subscales. Current cut scores are at levels of EBD risk that are more informative for males as compared to females. Such a discrepancy could potentially lead to over-classifying males as having higher risk for EBD than females. This finding aligns with current research suggesting that males are more frequently classified as at risk for EBD than females (Coutinho, & Oswald, 2005; Young, Sabbah, Young, Reiser, & Richardson, 2010).

If, however, the scale or scale scoring is revised as suggested according to the response option changes indicated by the NRM to introduce more measurement precision, a high risk cut score of 3 could be recommended for the internalizing scale, and a high risk cut score of 7 for females and 5 for males on the externalizing cut score. The difference in measurement precision of the subscales by gender could be evidence of middle school teachers who may be more likely to endorse males as exhibiting externalizing behaviors as opposed to females as suggested in previous research or due to an actual difference between males and females in EBD risk (Young, Sabbah, Young, Reiser, & Richardson, 2010). Either way, revised cut scores modeled in this way would provide greater measurement precision in identifying those most likely to be at risk for EBD. Additionally, evidence is provided in these analyses that using only one cut score for both males and females may oversimplify the relationship between gender and EBD risk as measured by the SRSS-IE.

Limitations and Implications for Future Research

The limitations of this research include the lack of diagnostic information corresponding with the screening data in this study. ROC curve analyses have the advantage of including data on the levels of sensitivity and specificity with which students are classified with EBD. IRT

analyses provide additional insight into the amount of measurement precision at given levels of theta; however, without subsequent diagnostic information about the students who were flagged at moderate or high risk of EBD, it is difficult to conclusively determine whether cut scores should be changed to match levels of peak measurement information.

Another limitation to this study is that the distribution of data for some of the items on the SRSS-IE is slightly skewed as is often the case when measuring psychological constructs (Preston & Reise, 2014). Some research suggests that the use of the NRM with non-normal data may bias parameter estimates, although sample size is a major factor in the degree of parameter estimation bias (i.e., large sample sizes, such as in this study, are less susceptible to bias; Preston & Reise, 2014).

Future research should include getting multiple ratings for each student from middle school teachers in various class periods throughout the day to ensure that ratings represent the full spectrum of students' behavior. More research with the recommendations to changes in item option configurations made in this study should include redesigning the instrument and retesting the scale with the newly configured items. Items should be reevaluated according to the procedures done in this research to ensure that the recommendations here provide the additional measurement precision and information of each item as described in this study. Lastly, researchers should check cut scores recommended here with subsequent identification of students with EBD to measure the sensitivity and specificity at which the SRSS-IE accurately identifies students with EBD.

Other research should include introducing questions about possible nesting effects which could be addressed using multilevel IRT. Ramsay Curve IRT (RC-IRT) methods could be used to address non-normal distributions of the data using newly developed software such as EQSIRT

(Multivariate Software, 2010) that can use RC-IRT with the NRM to estimate parameters (Preston & Reise, 2014). Lastly, the degree of DIF on these items on the SRSS-IE should be examined using other demographic variables (e.g., ethnic group) that have historically shown some differences in ratings of mental health (Epstein, et al., 2005; Tyson, 2004).

Conclusion

In conclusion, identifying some degree of DIF and DSF effects on each of the items on the SRSS-IE could help suggest the possibility of revision or deletion of items that are not functioning as intended. Additionally, examining specifically whether differential functioning occurs at the step levels could help to identify and subsequently remove sources of bias at particular trait levels. An examination of gender differences in cut scores can also provide additional information to ensure accurate identification of both males and females at risk for EBD. Such revisions to the scale and to the cut scores would ensure that the SRSS-IE is functioning well as a screening tool that is critical to identifying middle school students who may be at risk for EBD. The better instruments such as the SRSS-IE can be at providing an unbiased measurement of EBD risk, the more likely such students are to get the help they need.

References

Achenbach, T. M., Howell, C. T., Quay, H. C., Conners, K., & Bates, J. E., (1991). National survey of problems and competencies among four- to sixteen-year-olds: Parents' reports for normative clinical samples. *Monographs of the Society for Research in Child Development, 56*, 1-130.

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA School-Age Forms & Profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families.

Cai, L. (2013). flexMIRT version 2: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

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*, 105-117.

Camilli, G., & Congdon, P. (1999). Application of a method of estimating DIF for polytomous test items. *Journal of Educational and Behavioral Statistics, 24*, 323-341.

Countinho, M. J., & Oswald, D. P. (2005). State variation in gender disproportionality in special education: Findings and recommendations. *Remedial and Special Education, 26*, 7-15.

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

DiStefano, C. & Morgan, G. (2011). Examining classification criteria: A comparison of three cut score methods. *Psychological Assessment, 23*, 354-363.

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

Epstein, J. N., Valencia, E. Y., Tonev, S. T., Arnold, L. E., Willoughby, M., Abikoff, H. B., & Hinshaw, S. P. (2005). The role of children's ethnicity in the relationship between teacher ratings and attention-deficit/hyperactivity disorder and observed classroom behavior. *Journal of Consulting and Clinical Psychology, 73*, 424-434.

Forness, S. R., Freeman, S. F., Paparella, T., Kauffman, J. M., & Walker, H. M. (2012). Special education implications of point and cumulative prevalence for children with emotional and behavioral disorders. *Journal of Emotional and Behavioral Disorders, 20*, 4-18.

Glover, T. A., & Albers, C. A. (2007). Considerations for evaluating universal screening assessments. *Journal of School Psychology, 45*, 117-135.

Hayden, E. P., & Mash, E. J. (2014). Child psychopathology: A developmental-systems perspective. In E. J. Mash & R. A. Barkley (Eds.), *Child psychopathology* (3rd ed., pp. 3-72). New York, NY: The Guilford Press.

Higa-McMillan, C. K., Francis, S. E., & Chorpita, B. F. (2014). Anxiety disorders. In E. J. Mash & R. A. Barkley (Eds.), *Child psychopathology* (3rd ed., pp. 345-428). New York, NY: The Guilford Press.

Hoffmann, M. L., Powlishta, K. K., & White, K. J. (2004). An examination of gender differences in adolescent adjustment: The effect of competence on gender role differences in symptoms of psychopathology. *Sex Roles, 50*, 795-801.

Kalberg, J. R., Lane, K. L., Driscoll, S., & Wehby, J. (2011). Systematic screening for emotional and behavioral disorders at the high school level: A formidable and necessary task. *Remedial and Special Education, 32*, 506-520.

Kamphaus, R. W., & Reynolds, C. R. (2007). *BASC™-2 Behavior and Emotional Screening System (BASC™-2 BESS)*. San Antonio, TX: Pearson.

Kauffman, J. M., & Landrum, T. J. (2009). *Characteristics of emotional and behavioral disorders of children and youth* (9th ed.). Columbus, OH: Merrill.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of *DSM-IV* disorders in the national comorbidity survey replication. *Arch Gen Psychiatry, 62*, 593-603.

Krueger, R. F., & Markon, K. E. (2006). Reinterpreting comorbidity: A model-based approach to understanding and classifying psychopathology. *Annual Review of Clinical Psychology, 2*, 111-133.

Lane, K. L., Bruhn, A. L., Eisner, S. L., & Kalberg, J. R. (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*, 211-224.

Lane, K. L., Kalberg, J. R., Lambert, W., Crnobori, M., & Bruhn, A. (2010). A comparison of systematic screening tools for emotional and behavioral disorders: A replication. *Journal of Emotional and Behavioral Disorders, 18*, 100-112.

Lane, K. L., Little, M. A., Casey, A. M., Lambert, W., Wehby, J. H., Weisenbach, J. L., & Phillips, A. (2009). A comparison of systematic screening tools for emotional and behavioral disorders: How do they compare? *Journal of Emotional and Behavior Disorders, 17*, 93-105.

Lane, K. L., Menzies, H. M., Oakes, W. P., & Kalberg, J. R. (2012). *Systematic screenings of behavior to support instruction: From preschool to high school*. New York, NY: The Guilford Press.

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*, 244-270.

Lane, K. L., Oakes, W. P., Cantwell, E. D., Menzies, H. M., Schatschneider, C., Lambert, W., & Common, E. A. (2016). Psychometric evidence of the SRSS-IE scores in middle and high schools. *Journal of Emotional and Behavioral Disorders,* 1-13. doi:10.1177/1063426616670862

Lane, K. L., Oakes, W. P., Cantwell, E. D., Schatschneider, C., Menzies, H., Crittenden, M., & Messenger, M. (2016). Student Risk Screening Scale for Internalizing and Externalizing Behaviors: Preliminary cut scores to support data-informed decision making in middle and high schools. *Behavioral Disorders, 42*, 271-284.

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 middle school level. *Assessment for Effective Intervention, 39*, 24-38.

Lane, K. L., Oakes, W. P., Ennis, R. P., Cox, M. L., Schatschneider, C., & Lambert, W. (2011). Additional evidence for the reliability and validity of the Student Risk Screening Scale at the high school level: A replication and extension. *Journal of Emotional and Behavioral Disorders, 21*, 97-115.

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*, 99-122.

Lane, K. L., Oakes, W. P., Lusk, M. E., Cantwell, E. D., Schatschneider, C. (2016). Screening for intensive intervention needs in secondary schools: Directions for the future. *Journal of Emotional and Behavioral Disorders, 24*, 159-172.

Lane, K. L., Oakes, W. P., Swogger, E. D., Schatschneider, C., Menzies, H. M., & Sanchez, J. (2015). Student Risk Screening Scale for Internalizing and Externalizing Behaviors: Preliminary cut scores to support data-informed decision making. *Behavioral Disorders, 40*, 159-170.

Lilienfeld, S. O. (2003). Comorbidity between and within childhood externalizing and internalizing disorders: Reflections and directions. *Journal of Abnormal Child Psychology, 31*, 285-291.

Liu, I-M, & Agresti, A. (1996). Mantel-Haenszel-type inference for cumulative odds ratios with a stratified ordinal response. *Biometrics, 52*, 1223-1234.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58*, 690-700.

McDermott, P. A. (1993). National standardization of uniform multisituational measures of child and adolescent behavior pathology. *Psychological Assessment, 5*, 413-424.

McDermott, P. A., & Weiss, R. V. (1995). A normative typology of healthy, subclinical, and clinical behavior styles among American children and adolescents. *Psychological Assessment, 5*, 162-170.

Moulton, S. E., & Young, E. L. (2016). *An examination of the psychometric properties of the SRSS-IE using the nominal response and generalized partial credit models*. Manuscript in preparation.

Multivariate Software (2010). EQSIRT: Item response theory software. Encino, CA: Author.

Nigg, J. T., & Barkley, R. A. (2014). Attention-deficit/hyperactivity disorder. In E. J. Mash & R. A. Barkley (Eds.), *Child psychopathology* (3rd ed., pp. 75-144). New York, NY: The Guilford Press.

Penfield, R. D. (2013). DIFAS: Differential item functioning analysis system (Version 5.0) [Software]. Available from http://www.education.miami.edu/facultysites/penfield/index.html

Penfield, R. D., & Algina, J. (2003). Applying the Liu-Agresti estimator of the cumulative common odds ratio to DIF detection in polytomous items. *Journal of Educational Measurement, 43*, 295-312.

Penfield, R. D., Gattamorta, K., & Childs, R. A. (2009). An NCME instructional module on using differential step functioning to refine the analysis of DIF in polytomous items. *Educational Measurement: Issues and Practice, 28*, 38-49.

Pomerantz, E. M., Altermatt, E. R., & Saxon, J. L. (2002). Making the grade but feeling distressed: Differences in academic performance and internal distress. *Journal of Educational Psychology, 94*, 396-404.

Preston, K. S. J. (2014a, April). *Advanced topics in IRT: Evaluating the effectiveness of each response option with the nominal response model*. PowerPoint presentation at the 94th annual convention of the Western Psychological Association, Portland, OR.

Preston, K. S. J. (2014b, April). *Advanced topics in IRT: Evaluating the effectiveness of each response option with the nominal response model*. PowerPoint presentation at the 94th annual convention of the Western Psychological Association, Portland, OR. Retrieved from http://hssfaculty.fullerton.edu/psychology/kpreston/Plotting.txt

Preston, K. S. J., & Reise, S. P. (2014) Estimating the nominal response model under nonnormal conditions. *Educational and Psychological Measurement, 74*, 377-399.

Preston, K. S. J., & Reise, S. P. (2015). Detecting faulty within-item category functioning with the nominal response model. In S. P. Reise & D. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 386-405). New York, NY: Routledge.

R Core Team. (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R- project.org/.

RStudio Team (2015). *RStudio: Integrated development for R*. RStudio, Inc., Boston, MA. Retrieved from http://www.rstudio.com/.

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

Tyson, E. H. (2004). Ethnic difference in using behavior rating scales to assess the mental health of children: A conceptual and psychometric critique. *Child Psychiatry and Human Development, 34*, 167-201.

Wagner, M., Friend, M., Bursuck, W. D., Kutash, K., Duchnowski, A. J., Sumi, W. C., & Epstein, M. H. (2006). Educating students with emotional disturbances: A national perspective on school programs and services. *Journal of Emotional and Behavioral Disorders, 14*, 12-30.

Walker, H. M., & Severson, H. H. (1992). *Systematic Screening for Behavior Disorders: Technical manual.* Longmont, CO: Sopris West.

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

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*, 225-235.

Yovanoff, P., & Squires, J. (2006). Determining cutoff scores on a developmental screening measure: Use of receiver operating characteristics and item response theory. *Journal of Early Intervention, 29*, 48-62.

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

Zwick, R., Thayer, D. T., & Mazzeo, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. *Applied Measurement in Education, 10*, 321-334.

APPENDIX A: flexMIRT Syntax

flexMIRT syntax for the seven SRSS-IE externalizing items modeled using the GPCM:

```
<Project>
  Title = "SRSS-IE";
  Description = "GPCM Analysis of the SRSS-IE";

<Options>
  Mode = Calibration;
      SE = SEM;
      smartSEM = Yes;
      SaveSCO = Yes;
      SavePRM = Yes;
      SaveDBG = Yes;
      SaveINF = Yes;
      SaveCOV = Yes;
      FisherInf = 81,4.0;
      Score = EAP;
      GOF = Extended;
      M2 = Full;
      FitNullModel = Yes;

<Groups>
  %Group1%
  File = "D:\SRSS-IE\SRSS Data\SRSS_All_Data_2015_flexmirt.dat";
  Varnames = v1-v12;
  Select = v1-v7;
  Missing = 9;
  N = 2122;
  Ncats(v1-v7) = 4;
  Model(v1-v7) = GPC(4);
 <Constraints>
```

flexMIRT syntax for the seven *original* SRSS-IE externalizing items modeled using the NRM:

```
<Project>
  Title = "SRSS-IE";
  Description = "IRT Analysis of the SRSS-IE";

<Options>
  Mode = Calibration;
      SE = SEM;
      smartSEM = Yes;
      SaveSCO = Yes;
      SavePRM = Yes;
      SaveDBG = Yes;
      SaveINF = Yes;
      SaveCOV = Yes;
      FisherInf = 81,4.0;
      Score = EAP;
      GOF = Extended;
      M2 = Full;
      FitNullModel = Yes;

<Groups>
  %Group1%
  File = "D:\SRSS-IE\SRSS Data\SRSS_All_Data_2015_flexmirt.dat";
  Varnames = v1-v12;
  Select = v1-v7;
  Missing = 9;
  N = 2122;
  Ncats(v1-v7) = 4;
  Model(v1-v7) = Nominal(4);
  Ta(v1-v7) =
      (0 0 0 0,
       1 0 0 0,
       1 1 0 0,
       1 1 1 0,
       1 1 1 1);
  Tc(v1-v7) = Trend;
<Constraints>
```

flexMIRT syntax for the seven *revised* SRSS-IE externalizing items modeled using the NRM:

```
<Project>
  Title = "SRSS-IE";
  Description = "IRT Analysis of the SRSS-IE EXT Items Revision 3";

<Options>
  Mode = Calibration;
      SE = SEM;
      smartSEM = Yes;
      SaveSCO = Yes;
      SavePRM = Yes;
      SaveDBG = Yes;
      SaveINF = Yes;
      SaveCOV = Yes;
      FisherInf = 81,4.0;
      Score = EAP;
      GOF = Extended;
      M2 = Full;
      FitNullModel = Yes;

<Groups>
  %Group1%
  File = "D:\SRSS-IE\SRSS Data\SRSS_All_Data_2015_flexmirt.dat";
  Varnames = v1-v12;
  Select = v1-v7;
  Missing = 9;
  Code(v1,v4,v5,v6)=
  (0,1,2,3),(0,1,1,1);
  Code(v2)=
  (0,1,2,3),(0,1,2,2);
  Code(v3,v7)=
  (0,1,2,3),(0,1,1,2);
  N = 2122;
  Ncats(v1,v4,v5,v6) = 2;
  Model(v1,v4,v5,v6) = Nominal(2);
  Ncats(v2,v3,v7)=3;
  Model(v2,v3,v7)=GPC(3);
  Ta(v1,v4,v5,v6) =
      (0,
       1);
  Tc(v1-v7) = Trend;
<Constraints>
```

flexMIRT syntax for the five SRSS-IE internalizing items modeled using the GPCM:

```
<Project>
  Title = "SRSS-IE";
  Description = "GPCM Analysis of the SRSS-IE";

<Options>
  Mode = Calibration;
      SE = SEM;
      smartSEM = Yes;
      SaveSCO = Yes;
      SavePRM = Yes;
      SaveDBG = Yes;
      SaveINF = Yes;
      SaveCOV = Yes;
      FisherInf = 81,4.0;
      Score = EAP;
      GOF = Extended;
      M2 = Full;
      FitNullModel = Yes;

<Groups>
  %Group1%
  File = "D:\SRSS-IE\SRSS Data\SRSS_All_Data_2015_flexmirt.dat";
  Varnames = v1-v12;
  Select = v8-v12;
  Missing = 9;
  N = 2122;
  Ncats(v8-v12) = 4;
  Model(v8-v12) = GPC(4);
 <Constraints>
```

flexMIRT syntax for the five *original* SRSS-IE internalizing items modeled using the NRM:

```
<Project>
  Title = "SRSS-IE";
  Description = "IRT Analysis of the SRSS-IE, Internalizing items";

<Options>
  Mode = Calibration;
      SE = SEM;
      smartSEM = Yes;
      SaveSCO = Yes;
      SavePRM = Yes;
      SaveDBG = Yes;
      SaveINF = Yes;
      SaveCOV = Yes;
      FisherInf = 81,4.0;
      Score = EAP;
      GOF = Extended;
      M2 = Full;
      FitNullModel = Yes;

<Groups>
  %Group1%
  File = "E:\SRSS-IE\SRSS Data\SRSS_All_Data_2015_flexmirt.dat";
  Varnames = v1-v12;
  Select = v8-v12;
  Missing = 9;
  N = 2122;
  Ncats(v8-v12) = 4;
  Model(v8-v12) = Nominal(4);
  Ta(v8-v12) =
      (0 0 0 0,
       1 0 0 0,
       1 1 0 0,
       1 1 1 0,
       1 1 1 1);
  Tc(v8-v12) = Trend;
<Constraints>
```

flexMIRT syntax for the five *revised* SRSS-IE internalizing items modeled using the NRM:

```
<Project>
  Title = "SRSS-IE";
  Description = "IRT Analysis of the SRSS-IE INT Items Revision 2";

<Options>
  Mode = Calibration;
      SE = SEM;
      smartSEM = Yes;
      SaveSCO = Yes;
      SavePRM = Yes;
      SaveDBG = Yes;
      SaveINF = Yes;
      SaveCOV = Yes;
      FisherInf = 81,4.0;
      Score = EAP;
      GOF = Extended;
      M2 = Full;
      FitNullModel = Yes;

<Groups>
  %Group1%
  File = "D:\SRSS-IE\SRSS Data\SRSS_All_Data_2015_flexmirt.dat";
  Varnames = v1-v12;
  Select = v8-v12;
  Missing = 9;
  Code(v9,v10)=
  (0,1,2,3),(0,1,1,1);
  Code(v8,v11,v12)=
  (0,1,2,3),(0,1,1,2);
  N = 2122;
  Ncats(v9,v10)=2;
  Model(v9,v10)=Nominal(2);
  Ncats(v8,v11,v12)=3;
  Model(v8,v11,v12)=GPC(3);
  Ta(v9,v10) =
    (0,
     1);
  Tc(v8-v12) = Trend;
<Constraints>
```

APPENDIX B: R Syntax

Sample R code for calculating CBD parameters and category intersection parameters, plotting CRCs and TIFs, and conducting Wald tests. Subsequent analysis in Rstudio required only changing the name of the data file and the flexMIRT syntax file name:

```
wd <- "D:/SRSS-IE/SRSS Data/flexmirt_Externalizing/"

flexname <- "flexMIRTconfig_EXT_Revision1"

source("http://hssfaculty.fullerton.edu/psychology/kpreston/Plotting.txt")

source("http://hssfaculty.fullerton.edu/psychology/kpreston/Wald.txt")
```

APPENDIX C: IRT Graphs

Original item CRCs with their respective item and category information curves (on the left) with revised item CRCs with their respective item and category information curves (on the right) for all 12 items on the SRSS-IE:

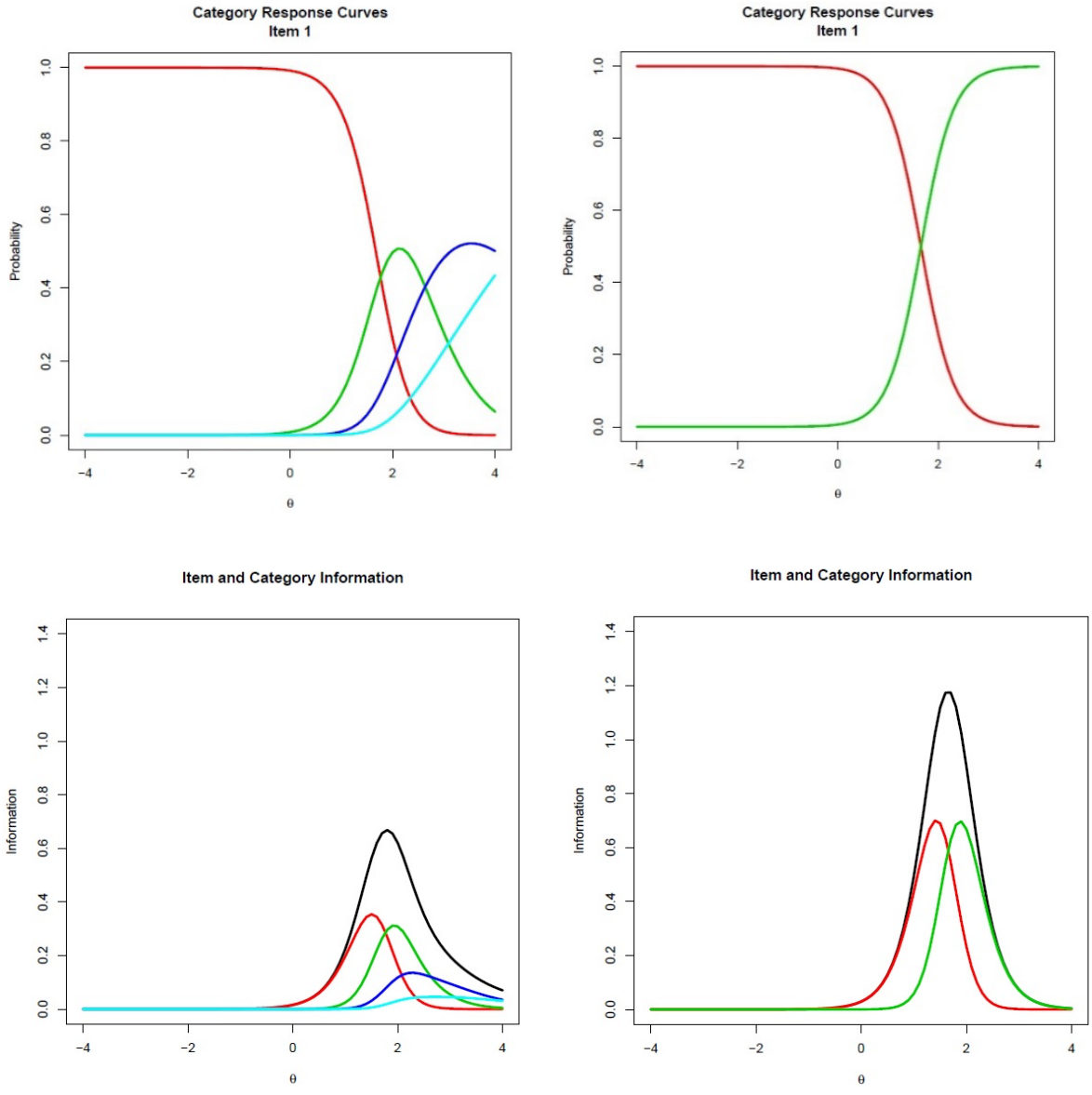

Figure 1. Stealing.

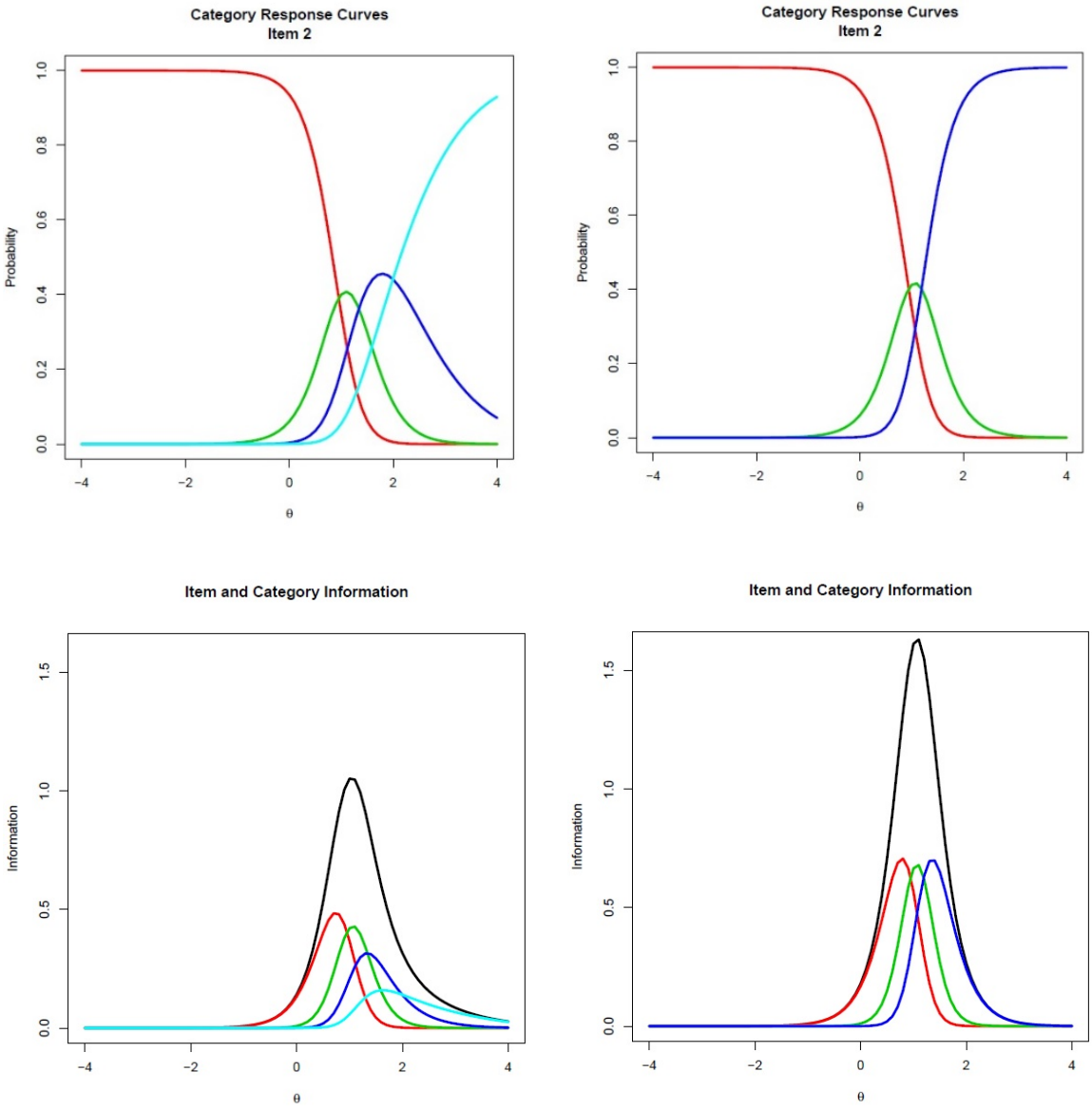

Figure 2. Lying, cheating, sneaking.

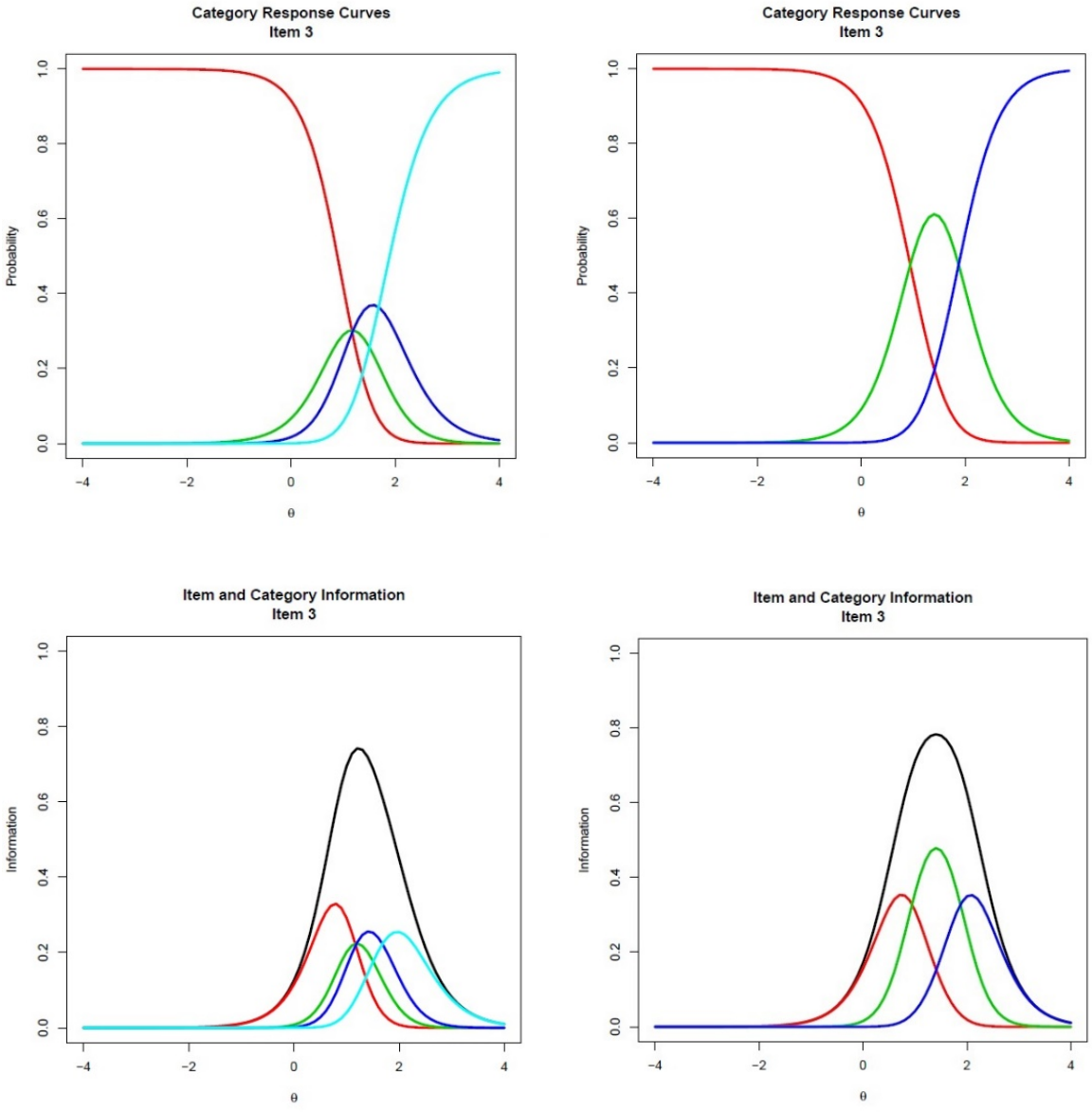

Figure 3. Behavior problems.

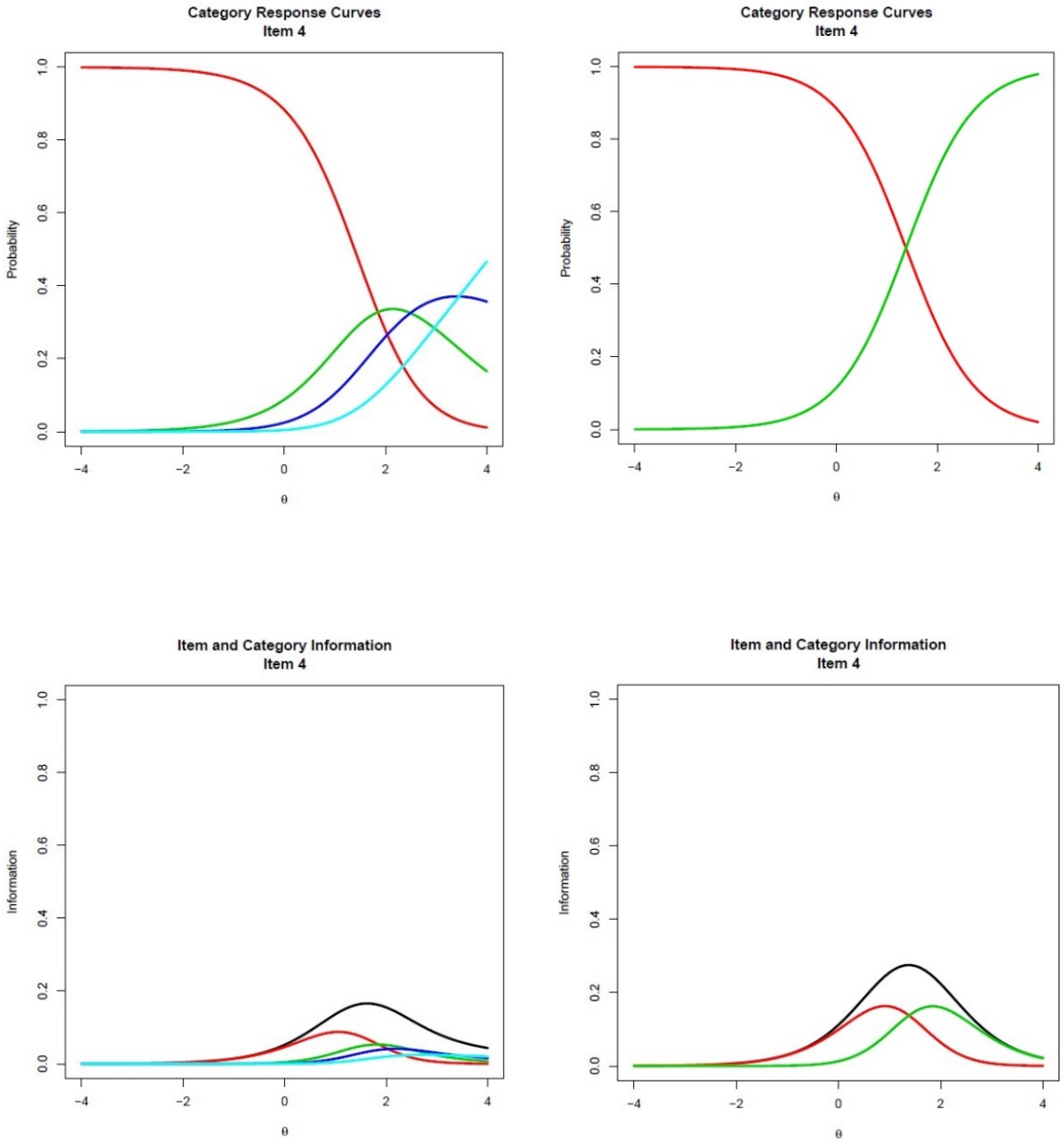

Figure 4. Peer rejection.

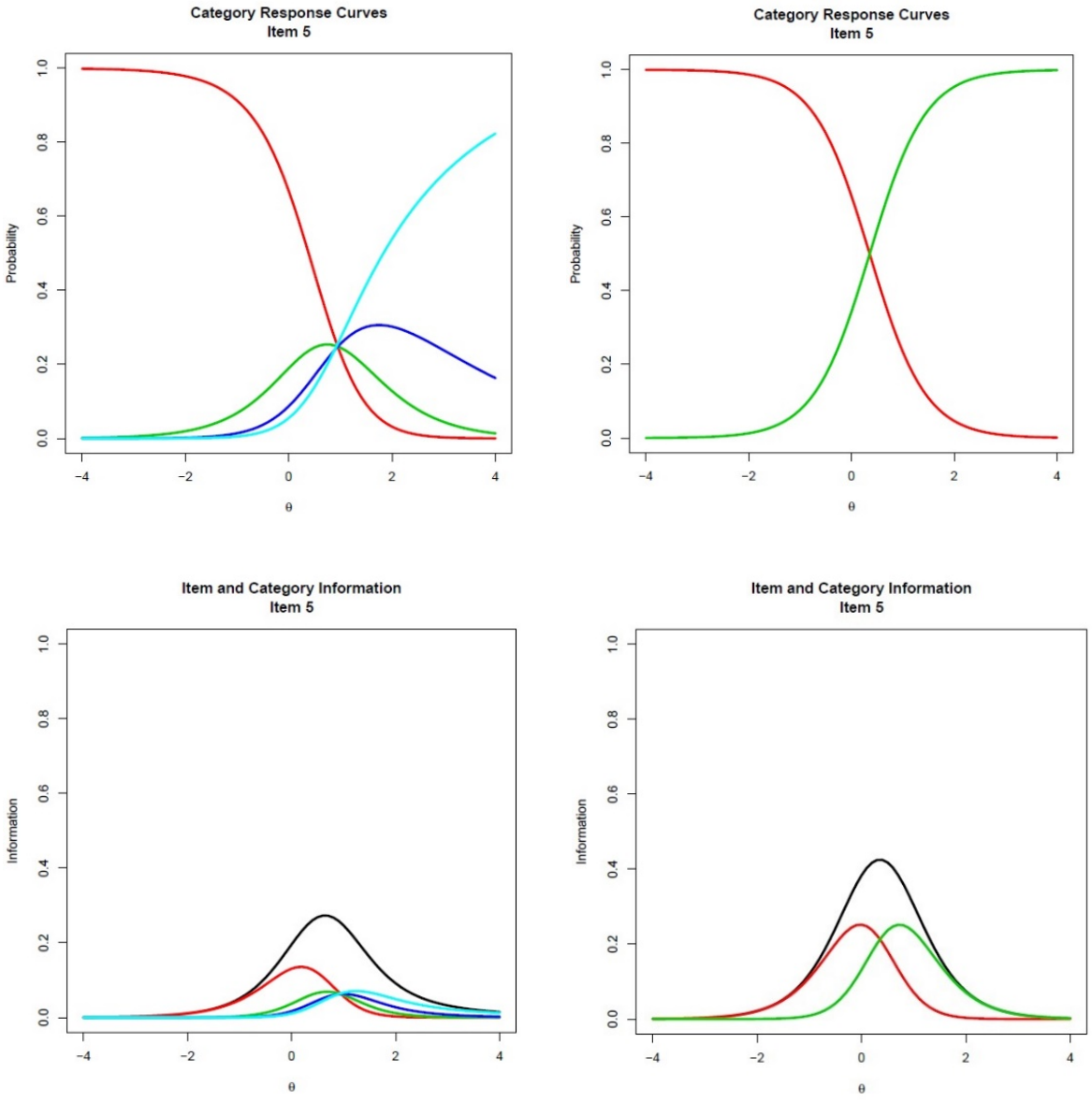

Figure 5. Low academic achievement.

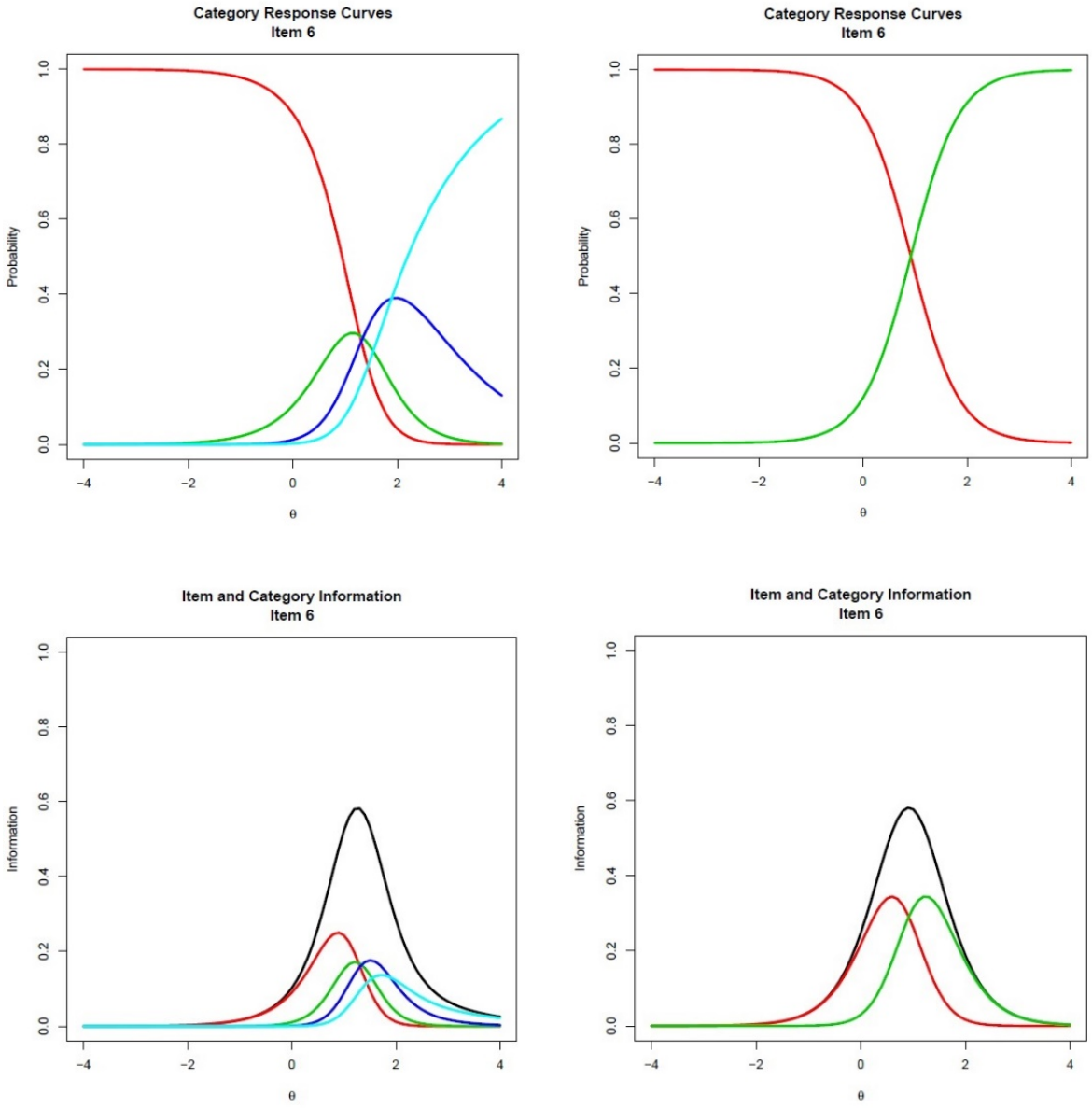

Figure 6. Negative attitude.

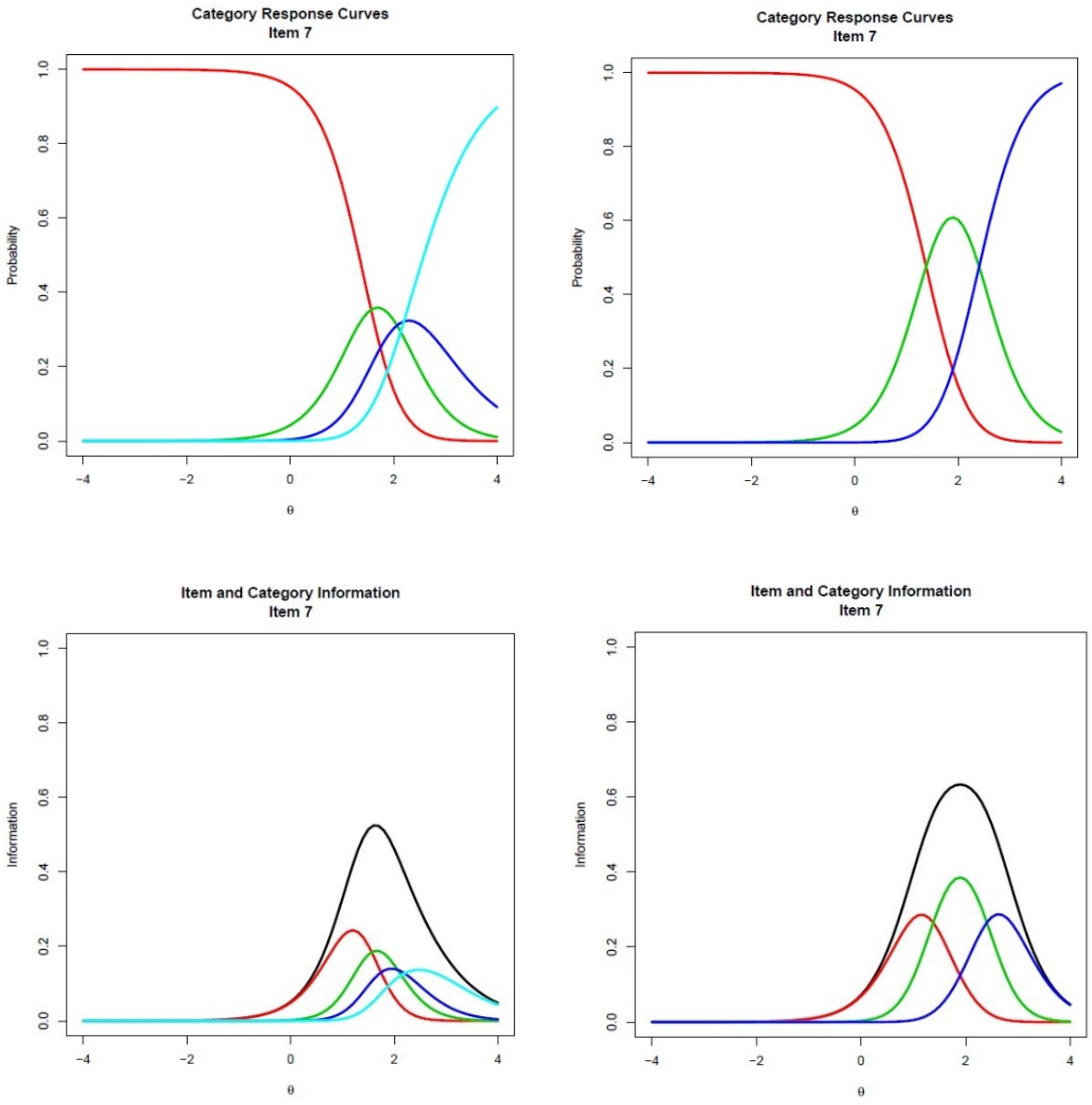

Figure 7. Aggressive behaviors.

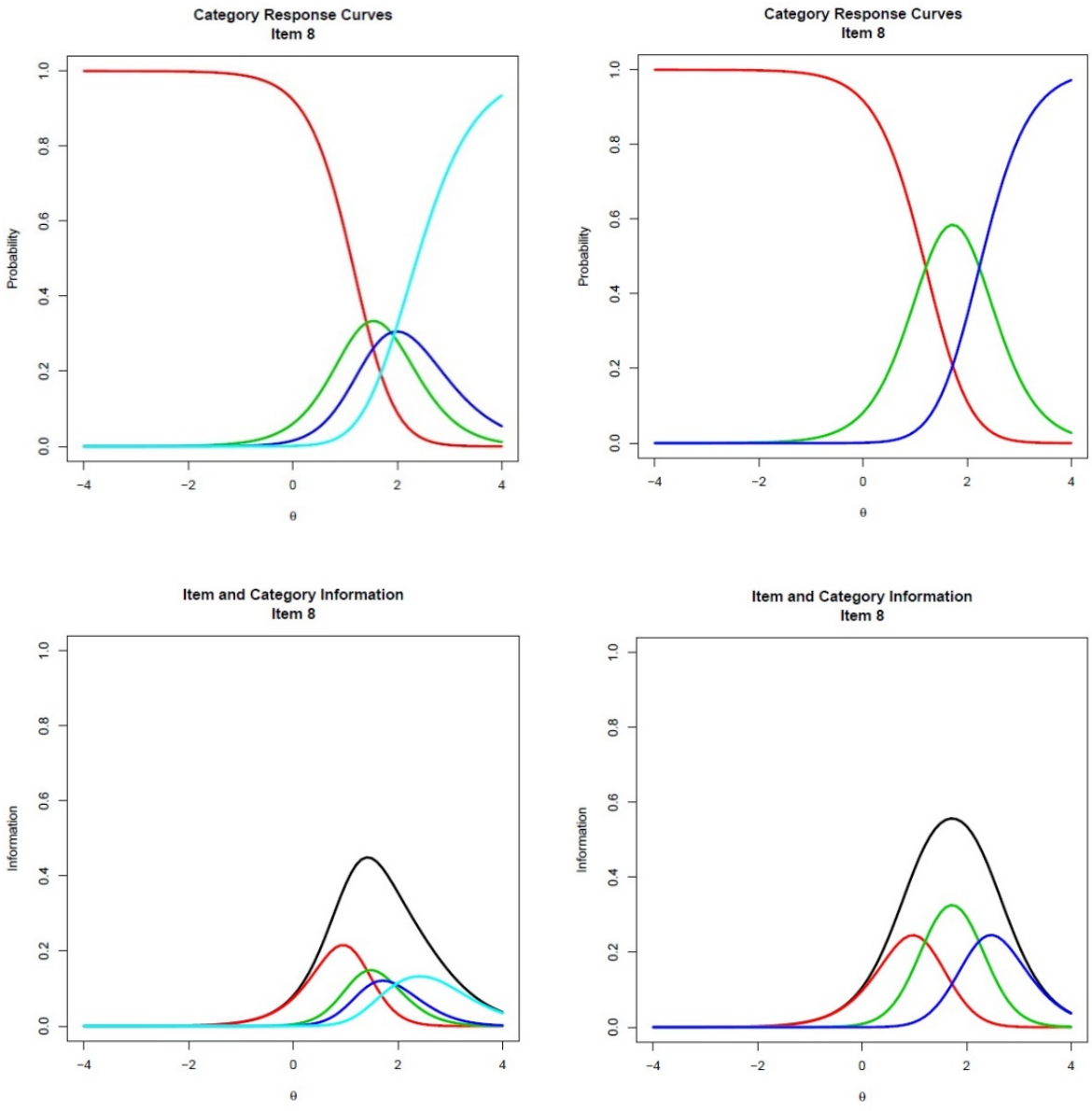

Figure 8. Emotionally flat.

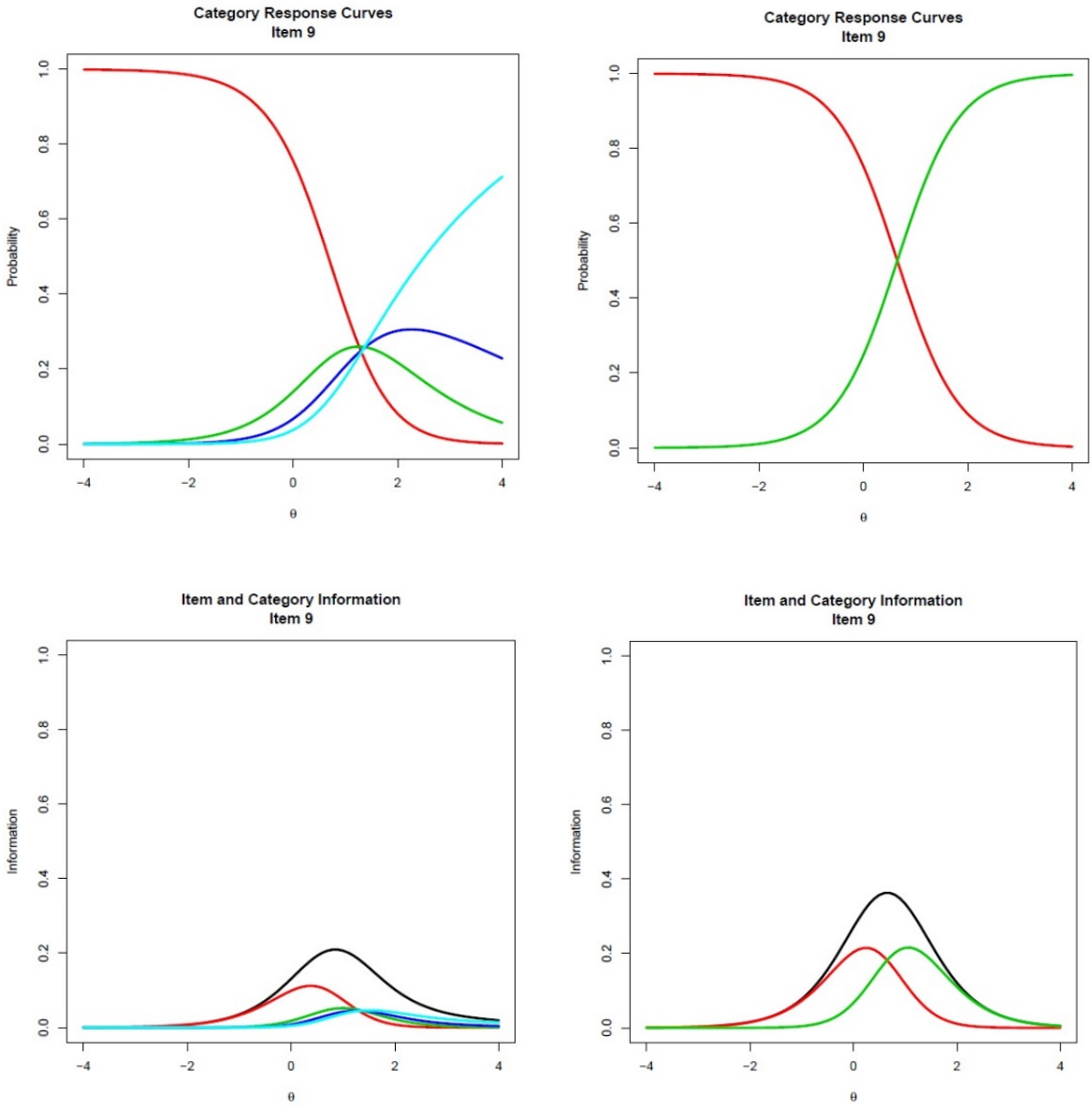

Figure 9. Shy, withdrawn.

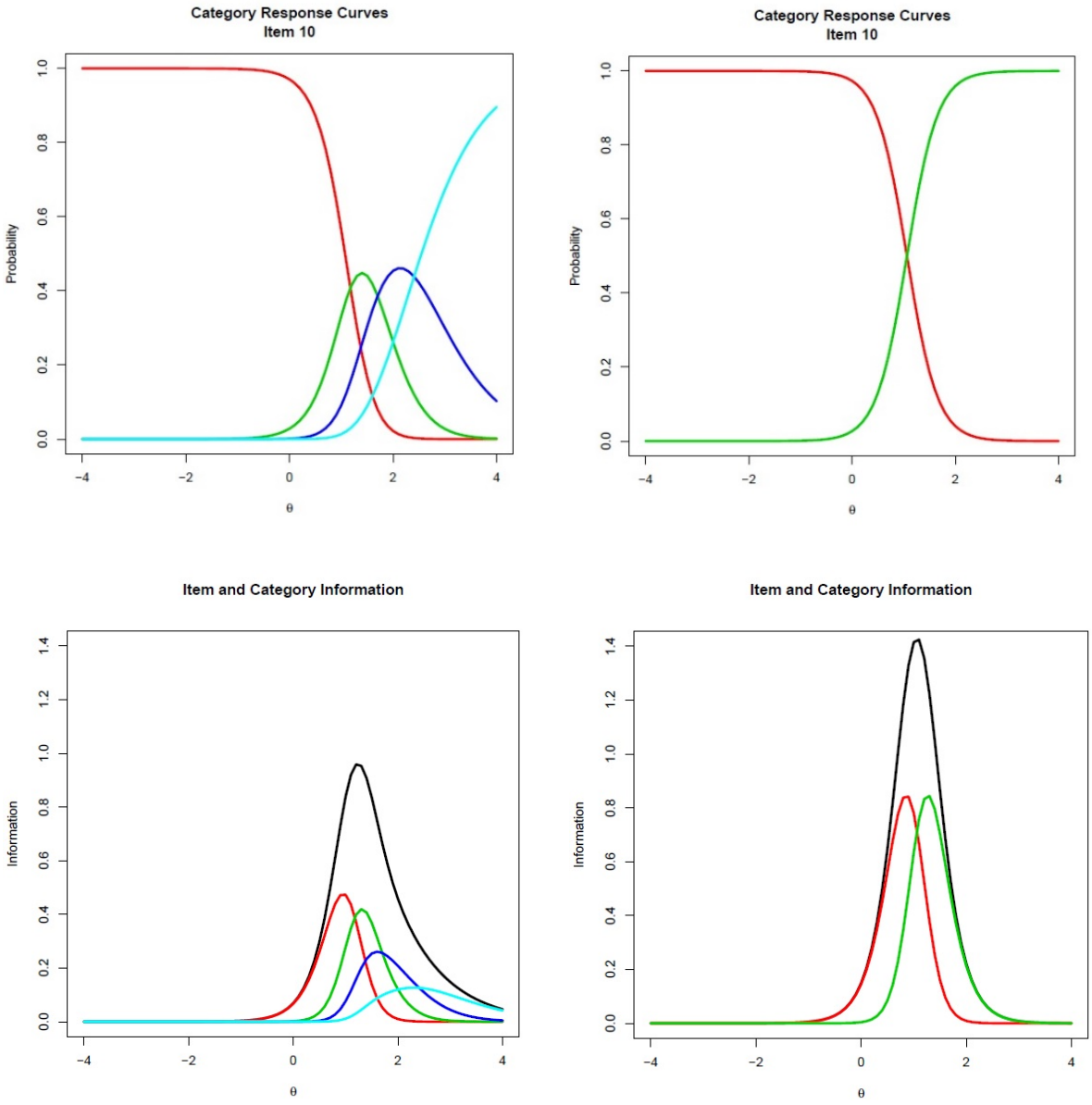

Figure 10. Sad, depressed.

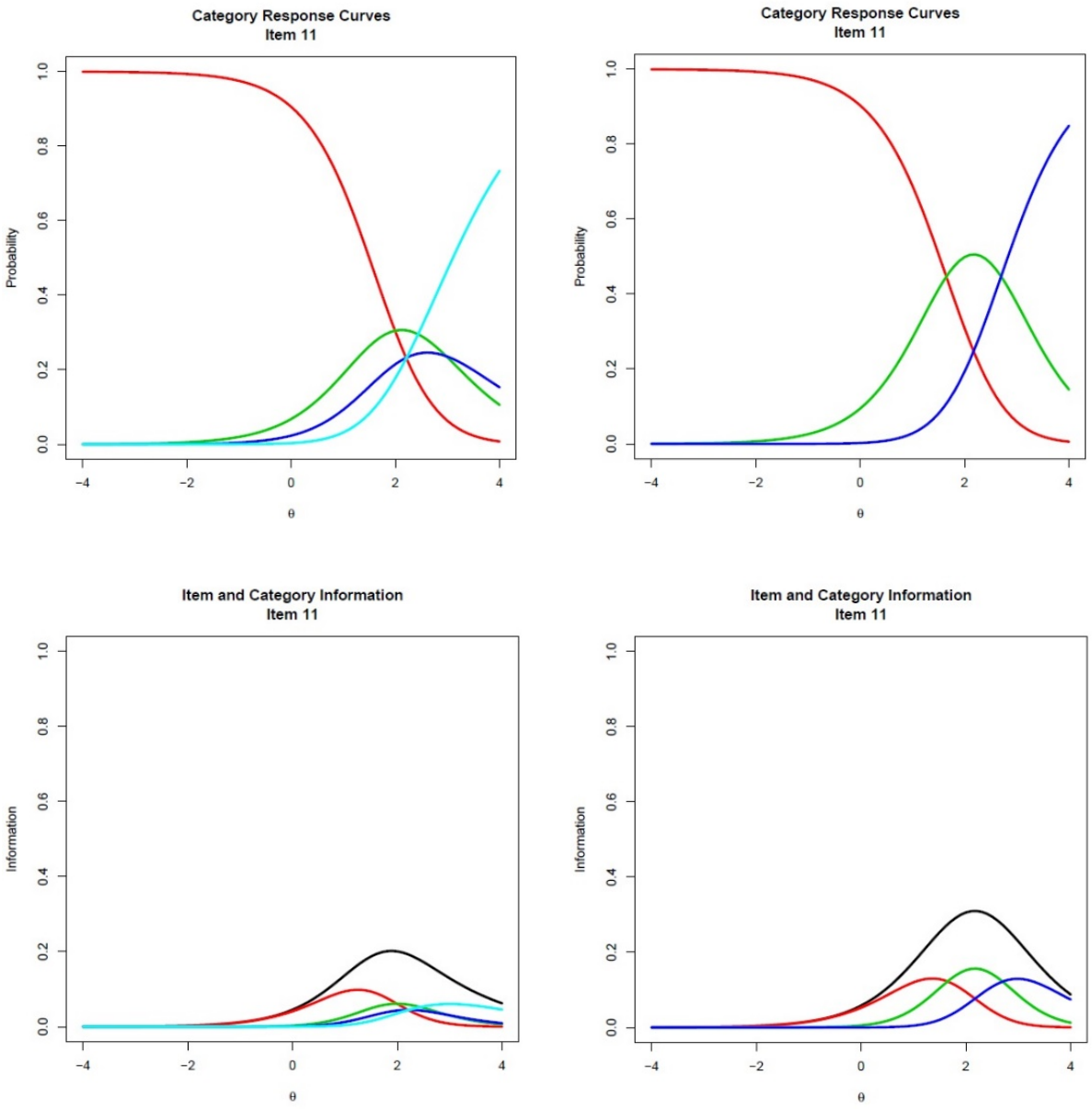

Figure 11. Anxious.

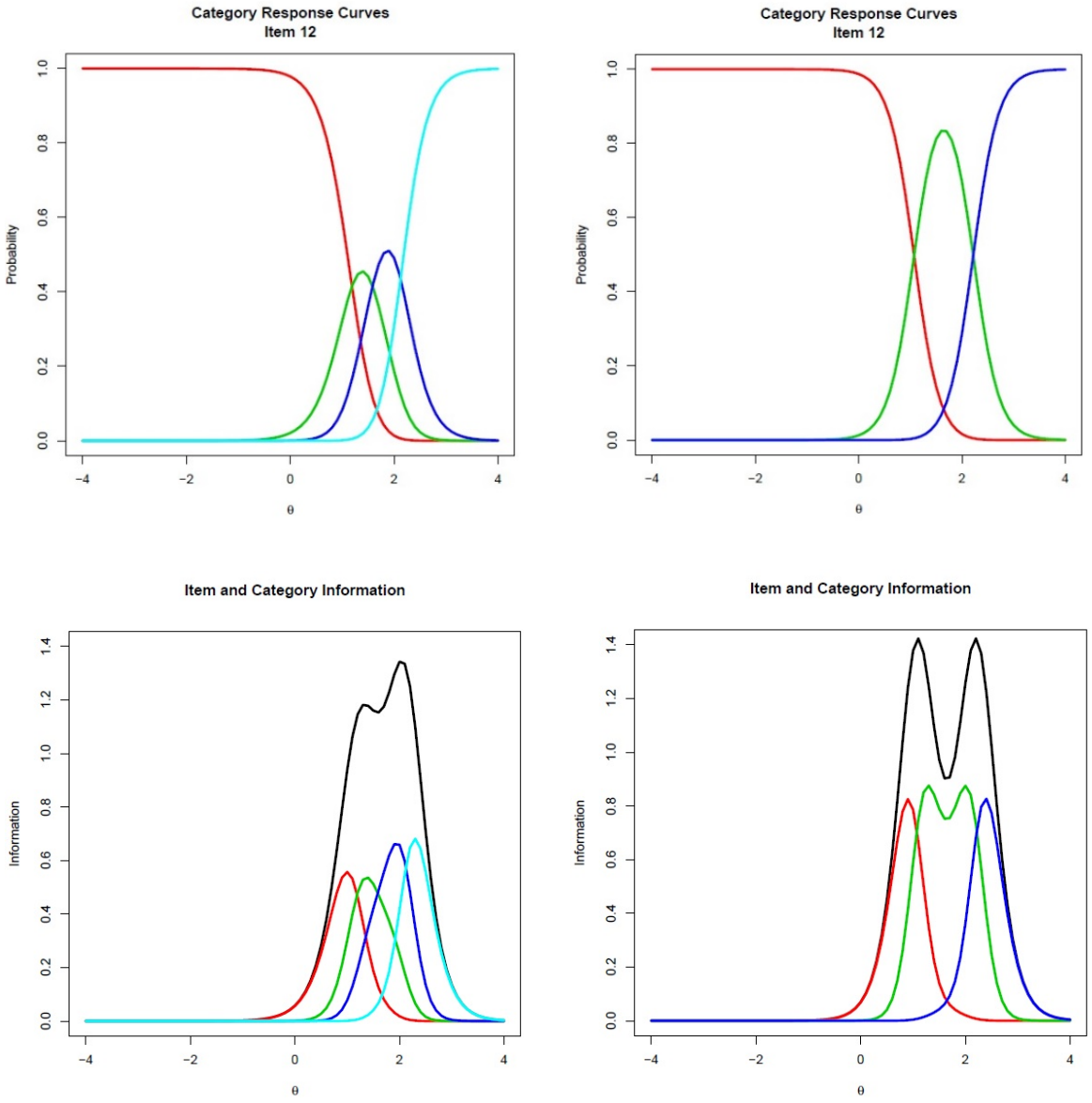

Figure 12. Lonely.