A MULTILEVEL MULTITRAIT-MULTIMETHOD ANALYSIS OF THE CHILD BEHAVIOR CHECKLIST

Marvin Powell, BSc., MSc.

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

August 2016

APPROVED:

Darrell M. Hull, Major Professor
Qi Chen, Committee Member
Anne Rinn, Committee Member
Bertina Combes, Committee Member and Interim Dean of the College of Education
Abbas Tashakkori, Chair of the Department of Educational Psychology
Victor Prybutok, Vice Provost of the Toulouse Graduate School

Powell, Marvin. *A Multilevel Multitrait-Multimethod Analysis of the Child Behavior Checklist*. Doctor of Philosophy (Educational Research – Research, Measurement and Statistics), August 2016, 90 pp., 6 tables, 4 figures, references, 149 titles.

Behavioral and emotional problems (BEPs) are known to affect children's ability to shape and maintain effective social relationships. BEPs are typically categorized into two main factors: internalizing and externalizing behaviors. Internalizing behaviors represent introverted problems, directed inwardly to the individual. While externalizing behavior patterns represent behaviors that are directed outwardly. Behaviors, emotions and thoughts are experienced by all people but on a continuum rather than in terms of absence versus presence of the behavior. The child behavior checklist (CBCL) is used to measure BEPs. The system of CBCL (parent form) measures also includes a teacher rating form and a youth self-report. Using 62 teachers and 311 students, the present study assessed convergent and discriminant validity using a correlated trait, correlated method minus one [CT-C(M-1)] model. The results showed low to moderate teacher-student agreement on the traits. To extend the theoretical structure of the teacher and self-report forms, the present study assessed the nested structure of the data using a multilevel model. Results revealed the nested structure of the data should not be ignored.

ACKNOWLEDGEMENTS

The Dissertation process is often a lonely road, filled with late nights and frustrations. My journey was no different; however, it was made more manageable with the assistance of God and several key people. Firstly, Dr. Darrell Hull, my committee chair, was instrumental in the successful navigation of the dissertation. Thank you sir for your keen attention to the little details that ensure a meaningful product. You are more than a committee chair, you are a mentor and a friend. I salute you. Thank you Dr. Bertina Combes for the mentoring and guiding my success throughout the program and this process. Dr. Anne Rinn, your positivity and willingness to facilitate my growth is the best, thank you. To Dr. Qi Chen, thank you for the expertise you provided to my understanding of the multilevel modeling. But I am more appreciative of you going beyond the "call" of duty. Dr. Abbas Tashakkori, I am thankful of your staunch advocacy of graduate students in the department. You believed in us. I must send a hearty thanks to Alecia Adams, Laura Coleman and Amber Brasher who made life so much easier at UNT.

My family has played a remarkable role in my life successes. Thank you mom, sis and Trace. Thank you for all your prayers. My friends from the Office of Research Consultancy thank you. To all my COE friends thank you for all the support and friendly faces. Dr. Sarah Ferguson, I could not have done this without you. We have done it!

Last but certainly not least, thank you Miss Lady for always providing the shoulder I needed when the process seemed unbearable. I share this success with you.

TABLE OF CONTENTS

Page

A MULTILEVEL MULTITRAIT-MULTIMETHOD ANALYSIS OF THE CHILD BEHAVIOR CHECKLIST

Introduction

Behavioral and emotional problems (BEPs) are known to affect children's ability to shape and maintain effective social relationships with peers, teachers, parents and family (Cullinan & Sabornie 2004; Gresham, Cook, Crews, & Kern, 2004; Landrum, Tankersley, & Kauffman, 2003). Negative behaviors are often disruptive leading to alienation from peers and denied learning opportunities from adults (Kauffman, 2001). Ultimately, unless BEP behaviors are curtailed, children's functionality in society will be impaired.

Children with BEPs are susceptible to negative educational and social outcomes at school. Nelson, Benner, Lane, and Smith (2004) examined K-12 public school students and found students with emotional and behavioral disorder had large academic achievement deficits. Students showed severe shortfalls in reading, written expression, and mathematics achievement even while controlling for age of onset of the problems (Nelson et al., 2004). Lane, Barton-Arwood, Nelson, and Wehby (2008) found similar results for both elementary and secondary students with behavior problems. BEP students typically score below the 25th percentile on reading, mathematics and written expression outcomes (Lane et al., 2008).

Interpersonal relationships, similar to academic outcomes, can be difficult for children with BEPs (Kauffman, 2001). BEPs in children tend to be effective predictors of problem behaviors in adulthood (Reef, van Meurs, Verhulst & van der Ende, 2010). In a 24-year longitudinal study, Reef et al. (2010) found that children who were reported (by parents) to have deviant behaviors tended to exhibit disruptive behaviors as adults; children with conduct disorder had mood and disruptive disorder at adulthood and; children with anxiety showed further anxiety

problems in middle adulthood. Moreland and Dumas (2008) similarly argued that behavior problems beginning at preschool age tend to be associated with life-long challenges. Gresham, et al. (2004) found social skills training effective in combatting several behavioral difficulties in children and youth. Early social skills interventions can prevent adult delinquency and substance abuse (Taylor & Biglan, 1998), decreased negative behavior, and increased prosocial skills with peers (Webster-Stratton, Reid, & Hammond, 2004).

However, with self-management interventions (self-monitoring, self-evaluation, strategy instruction and goal setting), students with BEPs can produce positive academic outcomes (Mooney, Ryan, Uhing, Reid, & Epstein, 2005). Students with emotional and behavioral problems show marked improvements in academic skills when introduced to these interventions. Similarly, teacher instruction and classroom management have shown to improve students' academic outcomes (Witt, Vanderheyden, & Gilbertson, 2004). Witt et al. argued that a deeper focus on instructional design (for example feedback and sequencing) would result in teachers having more success redirecting students' academic and behavioral outcomes. Teachers have identified that the school process can also be successfully navigated if students with behavior problems enhance self-control or cooperation skills (Lane, Pierson, & Givner, 2004; Polsgrove & Smith, 2004).

Measuring Behavioral and Emotional Problems in Children

Behavioral and emotional problems (BEPs) are characterized into two factors that conceptualize deviant behaviors (Coleman & Webber, 2002). The two factors are labeled internalizing and externalizing behaviors. Internalizing behaviors represent introverted problems, directed inwardly to the individual. Children exhibiting internalizing problems tend to present subtle markers and typically go unnoticed (Gresham & Kern, 2004). Examples of internalizing

behaviors include worries, social withdrawal, depression, anxiety, and obsessive-compulsive behaviors. On the other hand, externalizing behavior patterns represent behaviors that are directed outwardly. Children with externalizing BEPs are likely to be more impulsive while internalizing children are more reflective. Examples of externalizing behaviors are disobedience, aggression, disruption, impulsivity, and temper tantrums.

The Child Behavior Checklist (CBCL) is a measure developed by *ASEBA.* The CBCL school-age forms were first created as the CBCL/4-18 (Achenbach, 1991; Achenbach & Edelbrock, 1983). The CBCL assesses the degree of informants' consistency to report behavioral and emotional problems in children/adolescents (Achenbach, McConaughy, & Howell, 1987). Since the inception of the CBCL/4-18, a number of validation studies of the measures have been conducted (Albrecht, Veerman, Damen, & Kroes, 2001; Greenbaum & Dedrick, 1998). The CBCL/4-18 was revised and is now known as CBCL/6-18 with a change in the age range (Achenbach & Rescorla, 2001).

The CBCL/4-18 and the revised CBCL/6-18 has a similar structure, consisting of eight syndrome scales. The eight syndrome scales are: Withdrawn, Somatic Complaints, Anxious/Depressed, Social Problems, Thought Problems, Attention Problems, Delinquent/Rule-Breaking Behavior, and Aggressive Behavior (Achenbach, 1991; Dedrick, Greenbaum, Friedman, Wetherington, & Knoff, 1997; Dedrick, Tan, & Marfo, 2008; De Groot, Koot, & Verhulst, 1994; Ivanova, Achenbach, Dumenci, et al., 2007). The eight syndrome scales were shown to have a higher-order factor structure with Delinquent/Rule-Breaking and Aggressive scales as indicators of *Externalizing Disorders* and Withdrawn, Somatic, and Anxious–Depressed scales as indicators of *Internalizing Disorders* (Achenbach, 1991; Greenbaum & Dedrick, 1998). The names of the grouping are due to within-self problems (internalizing) and

for externalizing problems concerning conflicts with others and with participants' expectations (Achenbach & Rescorla, 2001).

The CBCL school-age forms are comprised of three components/forms: CBCL/6-18, completed by parents or surrogates, Youth Self-Report (YSR), and Teacher's Report Form (TRF). The CBCL is a multicultural problems assessment instrument with normative samples for over 40 societies (Achenbach & Rescorla, 2007). The syndromes are named to represent the description of the problems and not as a diagnosis (Achenbach, 1991; McConaughy, 1993). However, Achenbach & Rescorla (2001) suggested that the DSM-IV diagnoses of similar descriptions are highly correlated with the syndromes.

Reports by Multiple Informants/Cross-informants

Collecting information from children across different situations and environments, ultimately using different informants, offers a more complete picture of children's behaviors. Behaviors should be reported consistently across situations and environments and behaviors should at least moderately correlate across informants (Achenbach, et al., 1987; Synhorst, Buckley, Reid, Epstein, & Ryser, 2005). Multiple methods/sources reporting behaviors provides "valuable information in the examination of…emotional and behavioral functioning in children and adolescents" (Renk & Phares, 2004, p. 240). Richardson and Day (2000) and Renk (2005) argued for the use of multiple informants along with self-reports to measure behaviors. The authors suggest children's behaviors tend to be specific to their situations and collecting data from multiple environments contributes to a holistic understanding of behavior. Additionally, raters may observe different types and severity of behaviors that otherwise would have not been self-reported.

Low correlations among multiple informants have been found when using the CBCL for: psychopathology (Reynolds & Kamphaus, 2004), adaptive behavior (Sparrow, Cichetti, & Balla, 2005), behavioral problems (Lee, Elliott, & Barbour, 1994) and social skills (Gresham, Elliott, Cook, Vance, & Kettler, 2010). However, Ladd & Kochenderfer-Ladd (2002), similar to Achenbach et al., 1987, found moderate agreement between youth report of behavior and adult reports. A meta-analysis conducted by Achenbach, Krukowski, Dumenci, and Ivanova (2005) found moderate to high correlations between self-report and informants on parallel instruments for substance abuse, internalizing and externalizing problems. Results from the meta-analysis revealed high rater agreement when externalizing problems are being measured. Additionally, the mean cross-informant correlations were not significantly different for externalizing and internalizing problems. Similarly, Achenbach (2006) found moderate cross-informant agreement for adults with aggression and rule-breaking problems and for adults with anxiety, depression and withdrawal problems.

Construct Validity and MTMM

Psychological concepts are difficult to measure explicitly; therefore, items are combined to create measures of the underlying constructs. This measurement difficulty drives researchers to develop methods where instruments are validated. The Standards for Educational and Psychological Testing (American Educational Research Association, 2014) highlighted this validation process as the extent to which a measure is supported by evidence and theory (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Messick (1989) defined validity as "the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores" (p. 13).

In this unified understanding of validity, Messick (1995) also highlighted six aspects of construct validity: (a) evidence of content relevance; (b) theoretical rationales for the observed consistencies; (c) fidelity of the scoring structure; (d) extent to which score properties and interpretations are generalizable across groups and settings; (e) convergent and discriminant evidence from multitrait-multimethod (MTMM) comparisons; and (f) value implications of score interpretation (p. 745). Establishing (a), (b) and (f) requires a deep theoretical agreement across members in the field. Psychometric evaluations are required to establish evidence of (c), (d) and (e). These aspects of validity address the magnitude of similarity of the item responses to the theoretical underpinning of the constructs.

MTMM matrices were first analyzed by Campbell and Fiske (1959) by eyeballing the correlation matrix following three rules. First, a measure is said to have convergent validity when measures of the same trait or ability have high correlations. Second, methods of collecting the data should discriminate different traits. This is evidence of discriminant validity. These correlation coefficients should be low, and lower in relation to the same-trait, different methods correlation (Kenny & Kashy, 1992). Finally, in order to assess method variability the correlation between same-method and different trait should be similar to the different-method, different-trait correlations.

Messick (1989) argued for the use of confirmatory factor analysis (CFA) in assessing construct validation. CFA, more generally latent variable or structural equation modeling (SEM), when being used to analyze MTMM data, separates measurement errors from individual differences due to method and trait effects. CFA MTMM models allow for testing relationships of trait and method variables with other latent variables. Additionally, all models can empirically test the underlying assumptions of the models. Subsequently, CFA models have been mostly

applied to analyzing MTMM data (Carretero-Dios, Eid, & Ruch, 2011; Cole, 1987; Eid et al., 2008; Flamer, 1983; Kenny & Kashy, 1992; Marsh & Bailey, 1991; Widaman, 1985).

Confirmatory Factor Analysis MTMM

Confirmatory factor analysis (CFA) is associated with testing theory (Messick, 1989). In the context of the present study, CFA assesses how well data fit a hypothesized measurement model (Tabachnick & Fidell, 2007). In a measurement model, specific items are loaded/weighted on theorized factors. This item-factor loading is referred to as the factor structure. In CFA-MTMM models the underlying factor structure is decomposed. Individual differences are partitioned with the identification of the influence of the measured trait, method, and the error (or unique) components. CFA-MTMM is predicated on the same assumptions outlined by Campbell and Fiske (1959) that "each measure loads on only one trait and one method factor, and that the covariances between trait and method factors are zero" (Maas, Lensvelt-Mulders, & Hox, 2009, p. 73).

Multiple CFA-MTMM models have been developed in the last three decades. Widaman (1985) posited a taxonomy of 16 CFA-MTMM models comparing four traits and four methods. Since then, other models have been developed (Eid, 2000; Eid, Lischetzke, Nussbeck, & Trierweiler, 2003; Kenny & Kashy, 1992; Marsh & Byrne, 1993; Marsh & Grayson, 1995; Marsh & Hocevar, 1988; Wothke, 1995). The correlated trait-correlated method (CT-CM; Marsh, 1989; Marsh & Grayson, 1995) and the correlated trait-correlated uniqueness model (CT-CU; Kenny, 1976) are the most used models (Eid et al., 2008).

An alternate model, correlated trait-correlated method minus one model [CT-C(*M*-1)] (Eid, 2000; Eid et al., 2003) was proposed, addressing shortcomings of the previous models. The CT-C(*M*-1) model is a unique variant of the CT-CM model. In the CT-C(*M*-1) modeling

approach, one method is selected as a reference method and contrasted with all other methods. This omitted method is also known as the comparison standard (Brown, 2006). The self-report method is often the most frequently used reference/comparison method. The main premise of the CT-C(*M*-1) model follows the notion that for each trait measured, the true score variables of the reference method are regressors in a latent regression where the dependent measures are the true score variables of the non-reference method(s) (Brown, 2006; Eid et al., 2003). The method effects then become residuals common to all measured variables by the same method.

The methods used to collect data should be the determining factor for selecting models to analyze MTMM data (Eid et al., 2008). One criterion in determining the type of model is whether the method is interchangeable or structurally different. Eid et al. (2008) outlined the differences between interchangeable and structurally different raters. Interchangeable methods (raters) involve randomly selecting participants to rate an individual. For example, several students were randomly selected from a class and asked to rate the teacher's performance. Alternatively, if a teacher completes a self-report form and a peer and supervisor also rate the teacher then the raters are deemed to be structurally different. The ASEBA forms are considered structurally different because the teachers and parents both rate the behaviors of the child, who also rate themselves. When comparing structurally different methods the CT-C(*M*-1) is the most appropriate model (Eid, Lischetzke & Nussbeck, 2006; Koch, Schultze, Burrus, Roberts, & Eid, 2015).

Assessing BEPs in the Caribbean and Belize

The Education Act in Belize, under the direction of the ministry, requires all children aged five to 14 years to receive an education. Secondary school enrollment for the academic year 2013-14 was approximately 21,000. Enrollment rates have steadily increased over the last decade

in Belize. However, “only 45 percent of secondary school-aged children attend school,” a value that is below the region’s average of 80 percent (Naslund-Hadley, Alonzo, & Martin, 2013, p. 4). One possible explanation for this low enrollment rate is limited access to secondary education for children in rural Belize (Naslund-Hadley et al., 2013).

The Belize government has become increasingly concerned about the rising number of criminal activities being committed by youth. In an attempt to reduce crime and recidivism in young people and young adults, the government implemented an island-wide school-based positive youth development (PYD) program and a juvenile rehabilitation program. The Ministry of Education has identified for some time the prevalence of behavior and emotional problems in schools as a barrier to effective learning and instruction.

The *ASEBA* CBCL forms have been widely used across multiple cultures (Ivanova, Achenbach, Dumenci, et al., 2007; Ivanova, Achenbach, Rescorla, Dumenci, Almqvist, Bathiche, et al., 2007; Ivanova, Achenbach, Rescorla, Dumenci, Almqvist, Bilenberg, et al., 2007). The TRF, CBCL and YSR measure behavioral and emotional problems in children from a multicultural perspective (Achenbach et al., 2008), yet little evidence exists as to the psychometric construct validity of the measure within or across cultures, or when cross-informant mechanisms are used to validate ratings. Presently, as far as it can be ascertained, there are only two studies that have analyzed the *ASEBA* CBCL forms using MTMM. The first study, Grigorenko, Geiser, Slobodskaya, and Francis (2010) assessed convergent and discriminant validity in a sample of Russian youth. The second (Gomez, Vance, and Gomez, 2014) conducted a similar study using Australian children and adolescents. The *ASEBA* has normed scores for 32 countries, with Jamaica and Puerto Rico being the only Caribbean nations included (Achenbach & Rescorla, 2007). The parent-teacher agreement of the *ASEBA* scales tends to be higher than

parent-child and teacher-child agreement (Achenbach, Dumenci, & Rescorla, 2002; Grigorenko, et al., 2010). Because data were collected under the auspices of the Ministry of Education in Belize, the only informants available were in-school youth and their teachers. Consequently, the present study will focus on the teacher-child agreement. The following questions were proposed for the present study:

1. Do teacher (TRF) and youth (YSR) forms differ for externalizing and internalizing psychopathological traits?
    a. What is the convergent and discriminant validity of the TRF and YSR?
2. When accounting for the multilevel structure where self-report measure is nested within the teacher report, can the factor structure be confirmed in a manner better fit to the data?
    a. What is the ratio of true to observed variance?

Method

To assess construct validity of the CBCL, the present study uses data from 48 of 60 secondary schools in Belize. The Ministry of Education, Youth and Sports governs and co-ordinates the education system in Belize. This body oversees schools from the six districts in the country: Belize, Cayo, Corozal, Orange Walk, Stann Creek and Toledo. Secondary education is provided mainly by schools with religious management (n = 23), but there are also government schools (n = 15), privately-managed schools (n = 6), and community-managed schools (n = 4; Ministry of Education, Youth and Sports, 2012). Belize is 8,800 square miles, has a population of 334,060, and is located in Central America neighboring Mexico, Guatemala and the Caribbean Sea (United Nations, 2013).

Participant Characteristics

Students enrolled in 1st and 2nd Forms (the equivalent of U.S. grades 8 and 9) and students from schools that offered technical/vocational education and their teachers participated in the present study.

Sampling Procedures

At the time of data collection, Belize had 60 secondary schools. Data collected for the present study were 86.7% (52) of the schools across the six districts. Ten experienced/former teachers from Belize were trained as enumerators to administer the YSR in late Spring 2015 to students in the schools. The enumerators received a common protocol for administration, and received instructions to provide a similar protocol to each of the teachers for completing the TRF, requesting that teachers select eight students with whom they were familiar that also had exhibited a variety of good and bad behaviors during the school year.

Sample

Sixty-two teachers (females = 54.8%) completed the TRF of the CBCL at the end of the school year, resulting in each teacher rating, on average, 4.6 students. Majority of the teachers were classroom teachers (69.4%), while 27.4% were counselors and two were administrators. Three hundred and eleven students (51.9% male) provided completed the YSF. Students ranged in age from 12 to 18 years (M = 14.88, SD = 1.5). Ethnicity of the students were as follows: Creole 26.9%, Garifuna 12.3%, Maya 5.6%, Metizo 41.4%, East Indian 3.1%, and 3.4% other ethnicity (7.4% did not report their ethnicity). Since the enumerators collected data at the conclusion of the school year, teachers knew the students they rated and their behaviors in class and on campus for approximately nine months.

Measure

Youth Self-Report

The current form of the YSR has been normed for children ages 11-18 (Appendix A). Demographic information and extra-curricular activities are requested on the first two pages. The next two pages contain 118 statements describing behaviors. Items are measured on a 3-point scale; "Not True", "Somewhat or Sometimes True" and "Very True or Often True". The items were coded so that higher scores reflected more of the construct. The number of items associated with each YSR syndromes are withdrawn (8 items), Somatic Complaints (10 items), Anxious/Depressed (13 items), Social Problems (11 items), Thought Problems (12 items), Attention Problems (9 items), Delinquent/Rule-Breaking Behavior (15 items), and Aggressive Behavior (17 items). Items are written in the first person (e.g. "I enjoy a good joke"). Cronbach's alpha values ranged from .64 to .84. Table 1 provides the items defining the syndromes.

Teacher Report Form

The current revised edition of the TRF is a rater form for teachers as the informant, and has been normed for children ages 6 -18. The TRF is completed by teachers or other school personnel (e.g. assistant teacher, principal, administrator or counsellor) describing children's functioning at school. The teacher form consists of 118 statements largely describing the same behaviors as in the YSR. Similarly, the items are measured on the "Not True", "Somewhat or Sometimes True" and "Very True or Often True" 3-point scale. Teachers rate how well the behavioral, emotional and social problems items applied to the target youth being rated (e.g. "Has difficulty learning"). Higher scores indicate the construct associated with the behaviors rated. The number of items associated with each YSR syndrome are (Table 1): Withdrawn (8 items), Somatic Complaints (9 items), Anxious/Depressed (16 items), Social Problems (11

items), Thought Problems (10 items), Attention Problems (26 items), Delinquent/Rule-Breaking Behavior (12 items), and Aggressive Behavior (20 items). Internal consistency coefficients (Cronbach's alpha) ranged between .81 and .93.

Procedure

The YSR was included in a larger battery of measures used in a program evaluation in Belize. The TRF was the only measure completed by the teachers of the students. The team of 10 enumerators and a survey supervisor administered the instruments. Enumerators (two per district) travelled to schools within the district, administered the instruments to students, and provided instructions for teachers in completing the TRF. The data collection procedure included the enumerators presenting a larger battery of instruments for the evaluation (including the YSR) during the class session.

Enumerators entered item-level responses on a spreadsheet. A double data entry procedure was implemented in which a separate enumerator duplicated the data entry on a new spreadsheet and the two separate data entries were compared to reduce data entry errors. The survey supervisor compiled the data files, flagging any entries that were discrepant between the two enumerators, going back to the original forms to correct any entry errors. The survey supervisor also conducted preliminary data cleaning by verifying that data values were within required range for each variable. A de-identified data file was then submitted for analysis. Subsequently, further data cleaning was performed by generating descriptive statistics, searching for outliers and extreme values.

Data Analysis

Statistical analyses were conducted using Mplus Version 6.12 (Muthèn & Muthèn, 2010). An exploratory analysis of the data revealed missing data of 25.9% or lower. Little's Missing

Completely at Random test suggested the data were missing at random [X^2 (27776) = 27382.67, $p > .05$]. Maximum likelihood (ML) algorithms addressed the missingness in data, through the modeling parameter estimates. Mplus models missingness for both categorical and continuous variables with non-ignorable missing data (Muthèn & Muthèn, 2010).

Assessing Convergent and Discriminant Validity

CT-C(*M*-1) models have been shown to effectively assess convergent and discriminant validity of the CBCL (Gomez et al., 2014; Grigorenko, et al., 2010). In the present study, a CT-C(*M*-1) model, as shown in Figure 2, was employed; where the self-report was used as the reference method and the teacher report the focal (non-reference) method. The self-report was the logical choice for the reference method because of interest in the deviation of teachers' ratings of behaviors from the self-report.

Rater effect may differ across traits. Moreover, to assess the rater-specific effects (difference between teachers and students) no less than two indicators were required for each trait-method combination. Furthermore, multiple indicator CTC(*M*-1) models were deemed more accurate than single indicator models (Eid, et al., 2003; Eid et al., 2008). In addition, there has been criticism of the ability of single indicator models to generalize across raters (Marsh, 1993). Therefore, for this analysis, split-half parcels were used for each syndrome, as recommended by Little, Cunningham, Shahar, and Widaman (2002). The observed scores used as the indicators for the CT-C(*M*-1) model were continuous and therefore a maximum likelihood estimation method in Mplus was applied to generate the results.

An alternate CT-C(*M*-1) model path diagram is shown in Figure 3. In this model, each syndrome is loaded onto the corresponding higher order factor and method factor, with the self-report being the reference method. The two directional arrow illustrates a correlation between the

Internalizing trait factor and the Externalizing trait factor. Consistency and method-specific coefficients were calculated from the standardized weights generated from the CT-C(*M*-1) model. Significant consistency coefficients provide evidence of construct validity of an indicator. For example, if a syndrome had a larger consistency coefficient than the method-specific coefficient, there was support for convergent validity. Discriminant validity of the trait was inferred by assessing the degree of correlations of the trait factors. Similarly, examination of the degree to which there was correlation of method factors was inferred as discriminant validity of methods. Low correlation coefficients provided weak support for discriminant validity. In both models, loadings of the first indicators were fixed to 1.00 to identify the scale of the latent reference factor and the method factor (see Figure 1 and Figure 2).

Root Mean square error of approximation (RMSEA), comparative fit index (CFI), and standardized root mean square residual (SRMR) were used to indicate model fit. Model fit was judged reasonable when CFI $\geq$ .95 and RMSEA $\leq$ .06 (Yu & Muthen, 2002) and SRMR $\leq$ .03. In deciding the most parsimonious model to retain when comparing alternative models, a cutoff of less than 0.01 change in incremental model fit indices was used (e.g. ΔTLI < -0.01), and a RMSEA increase of less than 0.015 (ΔRMSEA < -0.015; Chen, 2007).

Extending Convergent and Discriminant Validity Evidence

In extending the convergent and discriminant validity of the TRF and the YSR, a multilevel correlated trait correlated method minus one model, ML-CT-C(*M*-1) was assessed as proposed by Koch et al. (2015); where student self-reports were nested within the teacher forms (Figure 4). In the ML-CT-C(*M*-1) model, using the within and between approach, analyses were based on the student level and the teacher level covariance matrix. The maximum likelihood estimator was used when analyzing ML-CT-C(*M*-1) model in Mplus (Muthen & Muthen, 2010).

RMSEA, CFI, $SRMR_{L1}$ and $SRMR_{L2}$ were used to determine model fit. Various variance decomposition coefficients, intraclass correlation (ICC) coefficients and reliability of the measures were determined by the ML-CT-C(*M*-1) model and reported to provide convergent and discriminant validity of the measures. The student-level ICC represents intra-cluster effects at the classroom level from the YSR. The teacher-level ICC represents intra-cluster effects at the classroom level as reported by the same teacher for a class of students on the TRF

Results

Descriptive Statistics

Descriptive statistics for both teachers and students on the eight syndromes (Withdrawn, Anxious/Depressed, Somatic Complaints, Social Problems, Thought Problems, Attention Problems, Rule-Breaking Behaviors and Aggressive Behaviors) are presented in Table 2. A close observation of the statistics reveals that students self-report behavioral and emotional problems more than their teachers, with the exception of Attention Problems. This pattern is the same for both split-halves of the syndromes as well as the total scores. The descriptive results also show higher variability in the TRF than in the YSR.

MTMM Results

The CT-C(*M*-1) model provides evidence of convergent validity for multiple informants when measuring traits. The trait-method combination for the present study is represented by 32 observed variables; 16 indicators (two for each of the eight syndromes), by two raters (teachers and students). A single model with all 32 indicators would have been comprised of approximately 500 parameters to be estimated which is larger than the sample size. To reduce the number of parameters estimated, three smaller models were specified.

The first model specified was for *internalizing* traits (Withdrawn, Anxious/Depressed, and Somatic Complaints). The second model consisted of *externalizing* traits (Rule-Breaking Behavior and Aggressive Behavior) and a third model contained the proposed second-order *problem behavior* traits (Social Problems, Thought Problems, and Attention Problems). The CT-C(*M*-1) models (Figure 2) were specified with the YSR as the reference group. All three models: Internalizing, $X^2(36) = 59.202$, $p < .01$, CFI = .988, RMSEA = .045 (90% CI .023, .065), SRMR = .024; Externalizing, $X^2(12) = 14.851$, $p > .05$, CFI = .999, RMSEA = .027 (90% CI .000, .066), SRMR = .017; and Problem Behaviors, $X^2(36) = 89.047$, $p < .0001$, CFI = .977, RMSEA = .068 (90% CI .050, .085), SRMR = .024, showed good fit to the data.

Convergent and Discriminant Validity

Table 3 shows the estimated means, factor loadings (trait and method), reliability, and variance decomposition coefficients (consistency and method specificity) for the eight syndromes of the TRF and YSR. The estimated unstandardized means are comparable to the means presented in Table 2. The standardized means allow comparison of raters' most and least frequently endorsed items for the syndromes. For example, both teachers and students tended to endorse items measuring Attention Problems and Withdrawn. Teachers were less likely to endorse Thought Problems items and students less likely to endorse items assessing Somatic Complaints.

Further assessment of the results in Table 3 reveals that on average, the method factor loadings are greater than the trait factor loadings. This indicates that YSR explains only a small proportion of the variance in the TRF. This interpretation is further supported by the consistency and method specificity coefficients. The consistency coefficients were relatively low across the syndromes, ranging from .00-.14. This provides evidence of the low agreement between the YSR

and TRF reports. Raters share the highest consistency on the Attention Problems trait (8%-14%) but practically no agreement on Withdrawn (0%-1%) and Thought Problems (1%). There are moderate to high method specificity coefficients for all syndromes, confirming the lack of agreement between the raters. The TRF, on average, shows higher reliability estimates than the YSR. Together, the consistency and method specificity coefficients provide adequate information to conclude low or no rater convergence.

The latent correlations among the trait factors are summarized in Table 4. These correlations represent the relationship between the syndromes as reported by the reference method (students). The correlations of the same trait indicator-specific factors (e.g., Withdrawn 1 and Withdrawn 2) are moderate in magnitude (r = .47 - .73). These correlations suggest the test halves relating to the same construct are heterogeneous. Homogeneous split halves measures should yield higher correlations (closer to 1.00; Grigorenko et al., 2010).

Inspection of trait correlations are more important because these correlations provide evidence of discriminant validity of the syndromes as reported by students. Low or no correlations between different traits reflect high degrees of discriminant validity. For example, correlations between different traits ranged between r = .25 and r = .55 for Internalizing problems. These correlations are moderate and indicate a low to moderate degree of discriminant validity. For the Externalizing problems, there is little discriminant validity support between Rule-Breaking Behavior and Aggressive Behavior (r = .62 - .74). The Problem Behavior traits do not discriminate with correlations ranging from .32 to .63 (Table 3).

Assessing method factors provides method effects generalizability across different traits with focus on the TRF. For the Internalizing traits, a high correlation (generalizability) was observed for Anxious/Depressed and Withdrawn (r = .76) and moderate correlations for

Anxious/Depressed and Somatic Complaints (r = .67), and Withdrawn and Somatic Complaints (r = .52). Generalizability for Externalizing traits could be deemed to be very high, with Rule-Breaking Behaviors and Aggressive Behaviors (r = .94). Finally, Problem Behaviors traits, Social Problems and Thought Problems (r = .86), Social Problems and Attention Problems (r = .72), and Thought Problems and Attention Problems (r = .70) are highly correlated. The moderate to high correlations suggest that, for example, teachers who rate students as possessing high aggressive behavior tend to rate the students high for rule-breaking behaviors relative to the students' evaluation.

ML-CTC(M-1)

The specified ML-CTC(*M*-1) model with trait and method factors, where the YSR was the reference method, shows adequate fit to the data, $X^2(50) = 139.66$, $p < .0001$, CFI = .945, RMSEA = .075, $SRMR_{L1}$ = .027, $SRMR_{L2}$ = .126. The specified model (Figure 4) contained two traits, Internalizing and Externalizing measured with two indicators, and two methods, the TRF and YSR. An alternative three-trait model (including the proposed Problem Behaviors trait) was assessed but the model did not converge. The results of the two-trait model are shown in Table 5.

Table 5 provides variance decomposition (consistency and method specificity) for both levels, reliability and intraclass correlation (ICC) coefficients for both Internalizing and Externalizing behavior traits. The reliability coefficients are high, ranging from .75 - .89, and greater for teachers, on both traits, than the student self-reports. This suggests problem traits are measured with greater accuracy (reliability) by the TRF than by the YSR. Additionally, teacher-level reports (.36-.42) produce higher ICC coefficients than student-level reports (.06-.08). This reveals that there is some level of dependency in the data; it is clustered because teachers rated different students.

The consistency coefficient at Level 1, regarding teacher reports is .04 for Internalizing and .01 for the Externalizing behavior trait. This means that there is higher convergence between teacher and student reports at the individual level (Level 1) for Internalizing behaviors than for Externalizing traits. However, these small coefficients suggest teachers' and students' evaluations were different at Level 1. Consistency coefficients are higher at Level 2 for both Internalizing and Externalizing traits in comparison to Level 1 consistency coefficients. This provides evidence of cluster level correlation between teacher and student reports, albeit low convergent validity.

The contrast to consistency coefficients is method specificity coefficients. Method specificity coefficients provide information regarding the proportion of the true variance of the teacher reports not shared with the student reports. The results in Table 5 reveal that at Level 1, method specificity coefficients ranged between .75 and .79. These high coefficients suggest poor teacher-student agreement on behavioral and emotional problems within classes. The coefficients are even higher at Level 2, .91 and .93, revealing lower teacher-student agreement across different classes.

Similar to the single level CTC(*M*-1) models, the multilevel models in the present study produce discriminant validity evidence by exploring latent trait factors and methods correlations at both levels. The latent trait factors (Internalizing and Externalizing) correlation at Level 1 is $r = .76$, $p < .001$ and $r = .80$, $p < .001$ at Level 2. These high and positive correlations indicate students perceived internalizing and externalizing traits as interrelated and did not differentiate between the two traits. The latent method factors correlation at Level 1 was $r = .58$, $p < .001$ and $r = .80$, $p < .001$ at Level 2. Method factors correlations demonstrate the degree to which

methods are generalizable across traits. The high method factor correlations suggest teachers rated students' traits (i.e., internalizing and externalizing behaviors) similarly.

Discussion

The aim of the present study was twofold. First, convergent and discriminant validity of the Child Behavior Checklist (CBCL) was assessed by applying a correlated trait correlated method minus one [CTC(*M*-1)] model. And second, a multilevel correlated trait correlated method minus one [ML-CTC(*M*-1)] model was used to further assess the convergent and discriminant validity of the multilevel structure of the data. In this nested structure, students were rated by their teachers on eight behavioral and emotional problems. These models provide important elements to MTMM data. For example, these models allow for the separation of measurement errors in assessing convergent and discriminant validity. More importantly, CTC(*M*-1) models provide variance decomposition into methods and trait factors for structurally different raters, which tend to be the data type in BEP research.

The results discussed focus on the low agreement between the teacher and student forms, the lack of trait discrimination, and the importance of considering the clustered nature of the *ASEBA* forms. The results from the single level CTC(*M*-1) model provide some supporting evidence of convergent validity of the TRF and YSR. Results were consistent with the findings of previous assessments of teacher-student ratings of problem behaviors (Gomez et al., 2014; Grigorenko et al., 2010). The findings of the present study showed there was more teacher-student convergence for Attention Problems than the other syndromes, albeit the consistency coefficient was low.

There are few studies conducted in the Caribbean assessing student-teacher rater agreement on the CBCL syndromes. Lambert, Knight, Taylor and Achenbach (1994) and

Lambert, Lyubansky and Achenbach (1998) found that a Jamaican sample had low teacher-student agreement on the eight syndromes. Low consistency across raters in the present sample is therefore not surprising; as culturally, Jamaica and Belize share similar historical and educational systems. It is conceivable that some student might suppress behaviors, specifically internalizing traits, to avoid deviant labels and punishments accompanying those behaviors. Therefore, the accuracy of teachers to observe BEPs in students are shadowed by the lack of students' outward expression of the behaviors.

Low rater convergence was more noticeable with high method specificity of the TRF. The teacher rating specificity is particularly high for Attention Problems and Somatic Complaints. With regard to Attention Problems, teachers are likely to be more aware of attention problems than students. If one considers a traditional classroom where teachers expect students to be attentive and focus so that they can maintain order, it is reasonable to assume that teachers would be keenly aware when students have lost focus. This higher level of awareness of student attention by teachers is a possible explanation for the lack of teacher-student agreement on Attention Problems. Similarly, items assessing Somatic Complaints are easily noticeable by teachers. For example, a teacher is likely to know of students' complaints of feeling dizzy, having a headache or stomachache. However, consistency was low for all syndromes. Such low agreement alludes to teachers and students perceiving the eight syndromes very differently.

Results revealed low to moderate discrimination among traits aligned with the second order latent variables: Internalizing, Externalizing, and Problem Behaviors. For internalizing traits (e.g. Withdrawn and Somatic Complaints, $r = .25$) there is evidence for discriminant validity. The other internalizing correlations (as well as externalizing and problem behavior traits) suggest the latent traits do not differentiate well. Raters in the present study perceive the

latent traits as more interrelated (and less separate) than suggested by the literature. This might be explained by cultural differences (Grigorenko et al., 2010). Individuals in Caribbean cultures tend not to differentiate internal and external motivations for a behavior; nor do they separate intrinsic and extrinsic behaviors. Once considered, behaviors are typically expressed externally and therefore, not separated as distinct (or different) behaviors. It is customary in Caribbean cultures to treat emotional and behavioral problems as internally motivated (Lambert et al., 2008).

The CBCL forms effectively examined BEPs in many populations and the present study supports its continued use in populations similar to Belize with the caveat of considering the cultural implications of interpreting the results. Previous studies also found low to moderate rater agreement suggesting the present results are not particularly unique. Two main theoretical perspectives have been purported for the low to moderate multi-informant agreement for youth behavioral problems: situation specific and bias reasons (Gomez et al., 2014). The situation specific hypothesis assumes rater differences is due to the environment (e.g., home, school etc.). The bias hypothesis refers to the variation in raters' perspective of the youth. Results from the present study align closer to the situation specific hypothesis as proposed by Achenbach et al. (1987). In order to determine the effectiveness of a measure, situation specificity must also include multiple cultures as much as multiple informants within cultures. Cultural considerations contribute to better understanding the assessment of BEPs in children and adolescents using CBCL forms.

The ML-CTC(*M*-1) model as proposed by Koch et al. (2015) assessed convergent and discriminant validity across the TRF and YSR. When the hierarchical (nested) structure of the data is ignored, the independence of the outcome is assumed. In a nested structure (continuing

with the example of students nested within classes), within-the-same-class students tend to be more similar than between-classes students by virtue of sharing the same teacher. These observations are therefore not independent because they depend on class membership. Additionally, when conducting CFA and its variants, ignoring hierarchical data structures introduces bias to chi-square fit statistics, parameter estimates and their standard error estimates (Julian, 2001; Koch, et al., 2015). The results of the present study show higher teacher intraclass correlations suggesting a lack of independence in the data. Consequently, ignoring the nested structure results in an inflation in the model fit estimates allowing for inaccurate conclusions regarding the data.

Limitations

Normative scores (*T* scores) have been recommended for ASEBA instruments (e.g. Gomez et al., 2014). However, no existing normative data is available for Belize, therefore the present study uses raw scores. Similar to Grigorenko et al. (2010), the aim of the present study was to assess convergent and discriminant validity and therefore was not concerned with the meaning of the scores.

Second, Hox and Maas (2001) proposed using at least 100 Level 2 units to conduct multilevel modeling. The present study utilized data from 62 teachers. Furthermore, more complex models had to be reduced to ensure the total number of parameters to be estimated did not exceed the sample size. A suitable alternative would have to have a larger number of Level 1 participants nested within teachers. However, teachers completed, on average, less than five student reports.

Finally, teachers interacted with students for approximately one hour per week for an average of nine months. The time teachers spent with students may be argued to be too limited to

form accurate opinions of their psychopathological symptoms. The *ASEBA* manual requires that the TRF be completed by individuals who have known the youth for at least two months (Achenbach & Rescorla, 2001). In the present study, an accumulation of one hour per week for nine months is more than equivalent to two months. Moreover, the teachers were asked to report how well they knew the student. Teachers reported knowing their students *moderately* to *very well* in 85% of cases.

Little et al. (2002) provide several techniques for building parcels. The present study used a random assignment procedure, creating two test halves. The test halves do not measure several syndromes with sufficient accuracy/reliability, suggesting a lack of parallel forms of the syndromes. Using an alternate parceling technique may yield additional findings.

Conclusion and Future Directions

Findings from the CTC(*M*-1) and the ML-CTC(*M*-1) models have important implications for continued use of the TRF and YSR measures in populations similar to Belize. The present study shows the CBCL reliably measures Internalizing and Externalizing traits in the Belize sample. And can be used with similar samples without factor structure reservations. However, there is little convergence between the TRF and YSR, and practically no discriminant validity, seemingly because raters view the traits as interrelated. Important information are provided regarding the teacher-student agreement once the nested structure is taken into consideration. When multi-informant measures are used in education, and teachers are the rater informants, it is essential to use analysis techniques that account for the clustered nature of the data.

The CBCL measures are useful forms to assess competences and behavioral problems. The forms cover an array of BEPs in children and adolescents. The measures can be used as a diagnostic tool or as a method of identifying and distinguishing the level of BEPs exhibited by

youth. However, the forms contain 119 items. The test length may be adequate as a diagnostic tool, but may prove lengthy for non-clinical use. There were no validation studies in the Caribbean setting and therefore, there was a lack of evidence whether the measures are effective in Belize. A detailed Item Response Theory analysis would provide information determining the functioning of each item in measuring the syndromes. This analysis would provide further psychometric evidence needed to determine construct validation as recommended by Messick (1995).

When applying the *ASEBA* measures, researchers and practitioners should use at least two of the three forms (Achenbach & Rescorla, 2001). Using the TRF provides a unique perspective of children psychopathological behaviors. Most importantly, researchers should always consider the nested structure of multi-informant measures used in schools and researchers interested in the psychometric properties of these instruments must consider the multi-level or hierarchical nature of these multi-informant tools that use teachers as the rater for several of their students. Including this form requires the use of a multilevel technique when assessing convergent and discriminant validity. ML-CTC(*M*-1) models allow for an accurate determination of construct validation evidence.

Table 1

Items Defining the YSR and TRF Syndromes Scales

Anxious Depressed	***Withdrawn Depressed***	***Somatic Complaints***	***Social Problems***	***Thought Problems***	***Attention Problems***
14. Cries a lot	5. Enjoys little	47. Nightmares[b]	11. Too dependent	9. Thoughts on mind	1. Acts young
29. Fears	42. Rather be alone	51. Feels dizzy	12. Lonely	18. Harms self	2. Odd noises[a]
30. Fears school	65. Refuses to talk	54. Overtired	25. Doesn't get along	40. Hears things	4. Fails to finish
31. Fears doing bad	69. Secretive	56a. Aches, pains	27. Jealous	46. Twitching	7. Brags[a]
32. Must be perfect	75. Shy, timid	56b. Headaches	34. Others out to get him	58. Picks skin	8. Can't concentrate
33. Feels involved	102. Lacks energy	56c. Nausea	36. Accident-prone	66. Repeats acts	10. Can't sit still
35. Feels worthless	103. Sad	56d. Eye problems	38. Gets teased	70. Sees things	13. Confused
45. Nervous, tense	111. Withdrawn	56e. Skin problem	48. Not liked	76. Sleeps less[b]	15. Fidgets[a]
50. Fearful, anxious		56f. Stomachaches	62. Clumsy	83. Stores things	17. Daydreams
52. Feels too guilty	***Rule Breaking Behavior***	56g. Vomiting	64. Prefers younger kids	84. Strange behavior	24. Disturbs others[a]
71. Self-conscious	2. Drinks alcohol[b]		79. Speech problem	85. Strange ideas	41. Impulsive
81. Hurt when criticized[a]	26. Lacks guilt			100. Trouble sleeping[b]	22. Difficulty with directions[a]
91. Talks or thinks of suicide	28. Breaks rules	***Aggressive Behavior***			49. Difficulty learning[a]
106. Anxious to please[a]	39. Bad friends	3. Argues a lot	76. Explosive[a]		53. Talks out of turn[a]
108. Afraid to make mistakes[a]	43. Lies, cheats	6. Defiant[a]	77. Easily frustrated[a]		60. Apathetic[a]
112. Worries	63. Prefers older kids	16. Mean to others	86. Stubborn, sullen		61. Poor schoolwork
	67. Runs away[b]	19. Demands attention	87. Mood changes		67. Disrupts discipline[a]
	72. Sets fires[b]	20. Destroys things	88. Sulks[a]		72. Messy work[a]
	81. Steals at home[b]	21. Destroys others' things	89. Suspicious		100. Fails to carry out tasks[a]
	82. Steals outside home	22. Disobey at home[b]	94. Teases a lot		73. Irresponsible[a]
	90. Swearing	23. Disobey at school	95. Temper		74. Show off
	96. Thinks of sex to much	37. Gets in fights	97. Threatens others		78. Inattentive
	98. Tardy[a]	57. Attacks people	104. Loud		80. Stare blankly[a]
	99. Uses tobacco	68. Screams a lot			92. Underachieving[a]
	101. Truant				93. Talks too much[a]
	105. Uses drugs				109. Whining[a]

Note. [a]Not on YSR; [b]Not on TRF

Table 2

Problem Syndromes Descriptive Statistics for Teachers and Students

Problems Scale	TRF		YSR	
	Mean	SD	Mean	SD
Withdrawn				
WD1	1.33	1.60	2.97	1.79
WD2	1.40	1.64	2.52	1.64
Total	2.74	2.99	5.49	2.96
Anxious/Depressed				
AD1	1.64	2.21	4.14	2.76
AD2	1.47	2.01	3.98	6.68
Total	2.94	3.77	8.12	4.98
Somatic Complaints				
SC1	1.04	1.74	2.54	1.96
SC2	0.66	1.24	2.68	2.04
Total	1.70	2.87	5.22	3.48
Social Problems				
SP1	1.57	1.92	3.50	2.47
SP2	1.31	1.67	2.75	1.97
Total	2.88	3.36	6.24	3.95
Thought Problems				
TP1	0.90	1.44	3.34	2.36
TP2	1.10	1.60	3.81	2.45
Total	2.00	2.82	7.14	4.30
Attention Problems				
AP1	4.72	4.78	3.03	2.07
AP2	4.99	5.13	2.90	1.93
Total	9.71	9.60	5.93	3.49
Rule-Breaking Behavior				
RB1	1.67	2.06	4.04	2.83
RB2	1.75	2.24	3.07	3.00
Total	3.41	4.07	7.11	5.30
Aggressive Behavior				
AB1	2.73	3.56	5.37	3.60
AB2	3.04	3.55	4.62	3.10
Total	5.77	6.81	9.99	6.24
Externalizing				
External 1	4.02	4.71	9.64	5.35
External 2	3.53	3.96	9.19	5.02
Total	9.18	10.50	17.09	10.97
Internalizing				
Internal 1	4.40	5.30	9.41	5.95
Internal 2	4.78	5.50	7.69	5.57
Total	7.38	8.16	18.83	9.68
Problem Behavior				
Problem1	7.19	7.01	9.86	5.82
Problem2	7.40	7.40	9.46	5.10
Total	14.59	14.15	19.32	10.37

Note. AD =Anxious/Depressed, WD = Withdrawn/Depressed, SC = Somatic Complaints, SP = Social Problems, TP = Thought Problems, AT = Attention Problems, RB = Rule Breaking Behaviors and, AG = Aggressive Behaviors.

Table 3

CT-C(M-1) Model for Teachers and Students: Intercepts, Factor Loadings, Consistency, Method Specificity, and Reliability of Observed Variables for the Eight Syndromes

		Factor Loadings			Variance Decomposition	
	Means (US/S)	Trait Factor (US/S)	Method Factor (US/S)	Reliability	Consistency	Method Specificity
			Withdrawn			
YSR 1	2.97/1.66	1.00/.75		.56		
YSR 2	2.52/1.54	.80/.65		.42		
TRF 1	1.33/0.83	-.01/-.01	1.00/.91	.83	.00	.83
TRF 2	1.40/0.86	.13/.11	.87/.77	.60	.01	.59
			Anxious/Depressed			
YSR 1	4.14/1.50	1.00/.82		.67		
YSR 2	3.98/1.49	.96/.82		.67		
TRF 1	1.64/0.74	.12/.13	1.00/.93	.88	.02	.86
TRF 2	1.47/0.73	.12/.13	.82/.84	.73	.02	.71
			Somatic Complaints			
YSR 1	2.54/1.30	1.00/.77		.59		
YSR 2	2.68/1.32	.90/.66		.44		
TRF 1	1.04/0.60	.25/.23	1.00/.92	.90	.05	.85
TRF 2	0.66/0.53	.13/.16	.68/.88	.80	.03	.77
			Social Problems			
YSR 1	3.50/1.42	1.00/.82		.67		
YSR 2	2.75/1.40	.68/.70		.49		
TRF 1	1.57/0.82	.13/.14	1.00/.87	.78	.02	.76
TRF 2	1.31/0.78	.15/.18	.83/.83	.72	.03	.69
			Thought Problems			
YSR 1	3.34/1.42	1.00/.77		.59		
YSR 2	3.81/1.56	1.06/.79		.62		
TRF 1	0.90/0.63	.10/.12	1.00/.76	.59	.01	.58
TRF 2	1.10/0.69	.09/.10	1.33/.92	.86	.01	.85
			Attention Problems			
YSR 1	3.03/1.47	1.00/.79		.62		
YSR 2	2.90/1.50	.78/.65		.42		
TRF 1	4.72/0.99	.86/.29	1.00/.87	.84	.08	.76
TRF 2	4.99/0.97	1.18/.37	1.09/.88	.91	.14	.77
			Rule-Breaking Behavior			
YSR 1	4.04/1.43	1.00/.85		.72		
YSR 2	3.07/1.03	.97/.77		.59		
TRF 1	1.67/0.81	.17/.19	1.00/.85	.76	.04	.72
TRF 2	1.75/0.78	.19/.20	1.14/.89	.83	.04	.79
			Aggressive Behavior			
YSR 1	5.37/1.49	1.00/.83		.69		
YSR 2	4.62/1.49	.91/.88		.77		
TRF 1	2.73/0.77	.21/.18	1.00/.90	.84	.03	.81
TRF 2	3.04/0.86	.17/.14	1.00/.90	.83	.02	.81

Note. US = Unstandardized; S = Standardized; YSR = Youth Self-Report; TRF = Teacher Report Form

Table 4

CT-C(M-1) Model: Latent Correlations Among Behavior Syndromes Reported by YSR

Indicators	WD1	WD2	AD1	AD2	SC1	SC2
			Internalizing Problems			
Withdrawn 1	-					
Withdrawn 2	.49	-				
Anxious/Depressed 1	.51	.47	-			
Anxious/Depressed 2	.55	.46	.67	-		
Somatic Complaints 1	.42	.25	.54	.47	-	
Somatic Complaints 2	.32	.32	.41	.45	.51	-

	RB1	RB2	AB1	AB2
		Externalizing Problems		
Rule-Breaking Behavior 1	-			
Rule-Breaking Behavior 2	.66	-		
Aggressive Behavior 1	.71	.62	-	
Aggressive Behavior 2	.74	.67	.73	-

	SP1	SP2	TP1	TP2	AP1	AP2
			Problem Behaviors			
Social Problems 1	-					
Social Problems 2	.58	-				
Thought Problems 1	.62	.57	-			
Thought Problems 2	.63	.50	.60	-		
Attention Problems 1	.60	.50	.50	.58	-	
Attention Problems 2	.50	.44	.32	.45	.52	-

Note. AD =Anxious/Depressed, WD = Withdrawn/Depressed, SC = Somatic Complaints, SP = Social Problems, TP = Thought Problems, AT = Attention Problems, RB = Rule Breaking Behaviors and, AG = Aggressive Behaviors.

Table 5

ML-CTC(M-1) Model with Unidimensional Latent Trait and Method Factors: Reliability, Intraclass Correlation, Consistency, and Method Specificity at Level 1 and Level 2

			Consistency		Method Specificity	
Rater	Reliability	ICCs	Level 1	Level 2	Level 1	Level 2
			Internalizing			
Students 1	.81	.08	.79	1.00		
Students 2	.81	.08	.79	1.00		
Teachers 1	.89	.36	.04	.09	.79	.93
Teachers 2	.89	.36	.04	.09	.79	.93
			Externalizing			
Students 1	.75	.06	.73	1.00		
Students 2	.75	.06	.73	1.00		
Teachers 1	.86	.42	.01	.09	.75	.91
Teachers 2	.86	.42	.01	.09	.75	.91

Note. ICC = intraclass correlation

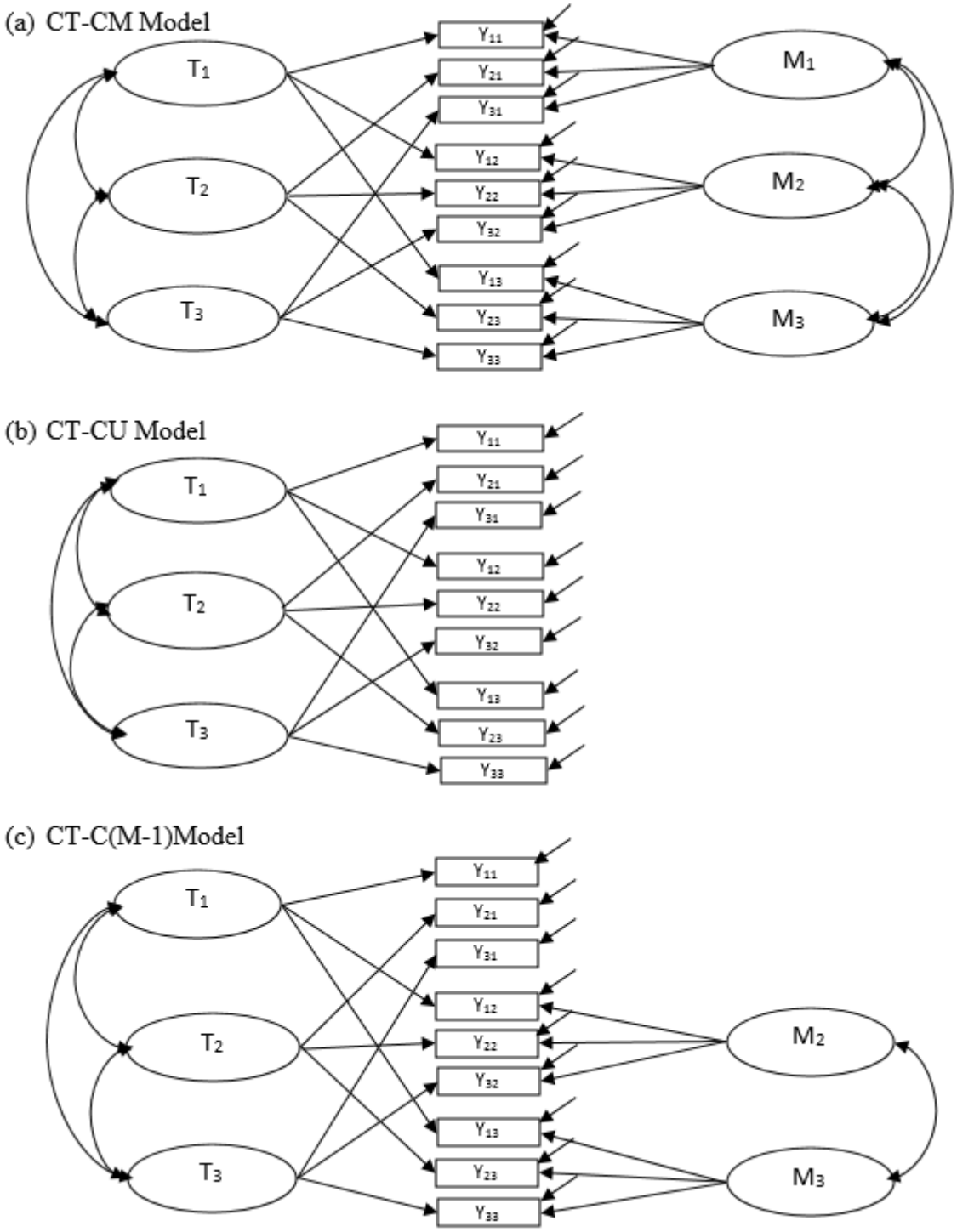

Figure 1. MTMM models: CT-CM = Correlated trait-correlated method model; CT-CU = correlated trait-correlated uniqueness model; CT-C(M-1) = Correlated trait-correlated method minus one model. Y_{jk} = observed variable; j = trait; k = method.

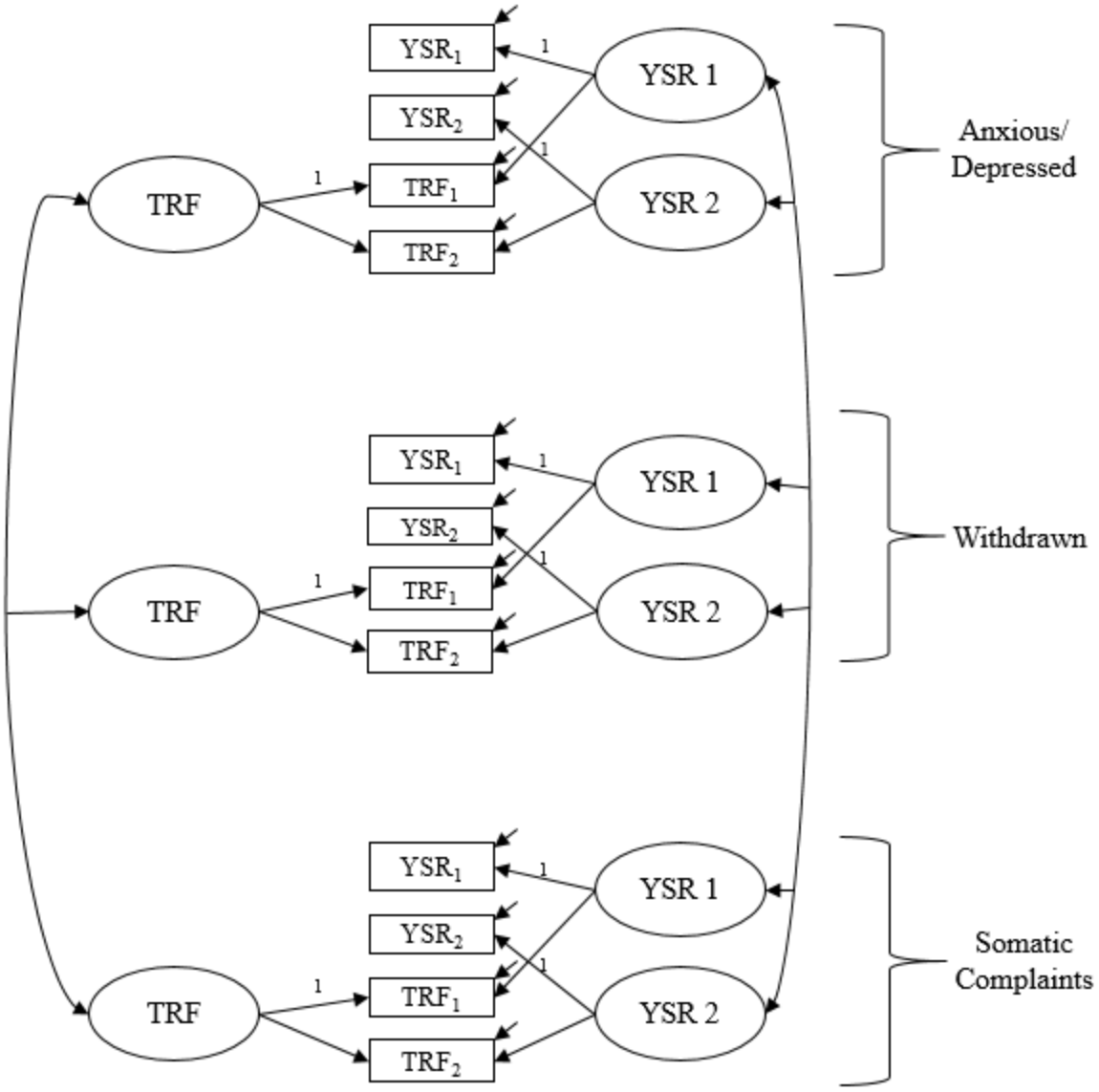

Figure 2. Correlated trait-correlated method minus one [CT-C(M-1)] model path diagram showing indicator-specific and trait-specific method factors for Internalizing traits (Anxious/Depressed, Withdrawn and Somatic Complaints). The indicators are represented by split-halves observed variables. The first loadings for each factor is fixed to 1 while the others are estimated freely. The Internalizing trait is only shown because of space, but Externalizing and Problem Behavior traits are represented similarly.

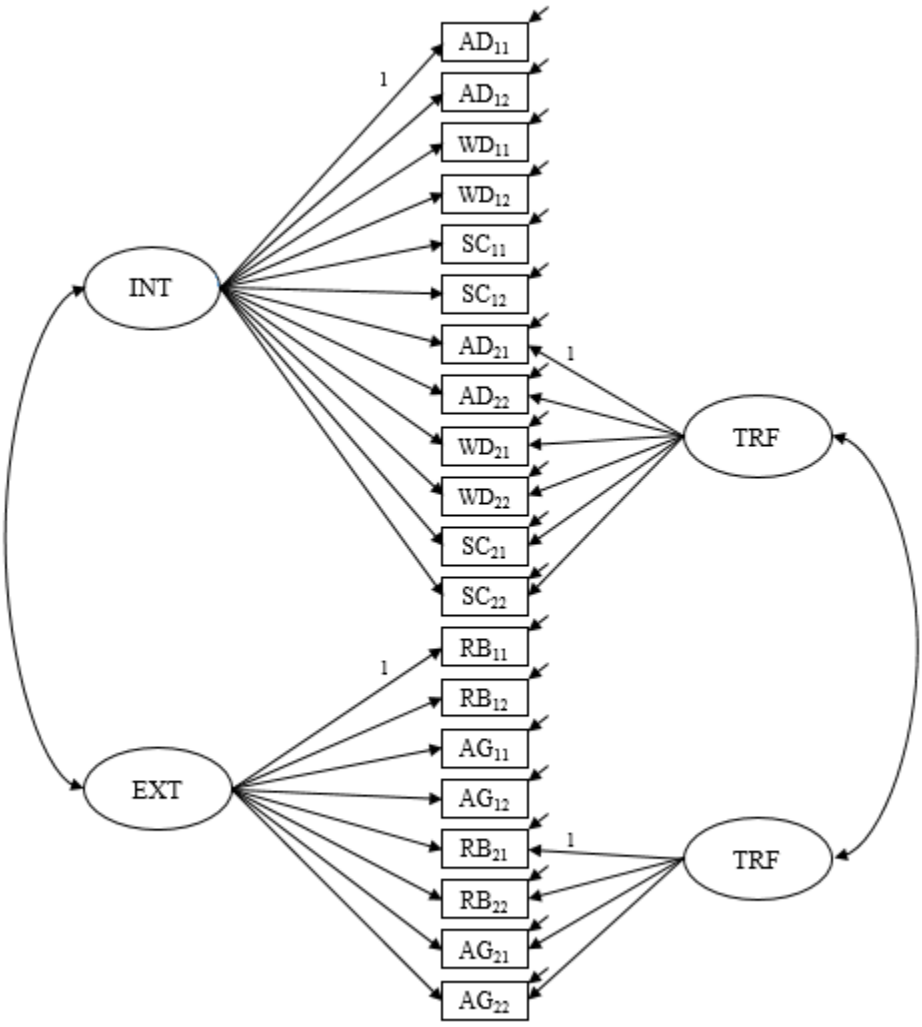

Figure 3. Hypothesized MTMM model: Correlated trait-correlated method minus one [CT-C(M-1)] model. AD =Anxious/Depressed, WD = Withdrawn/Depressed, SC = Somatic Complaints, SP = Social Problems, TP = Thought Problems, AT = Attention Problems, RB = Rule Breaking Behaviors and, AG = Aggressive Behaviors

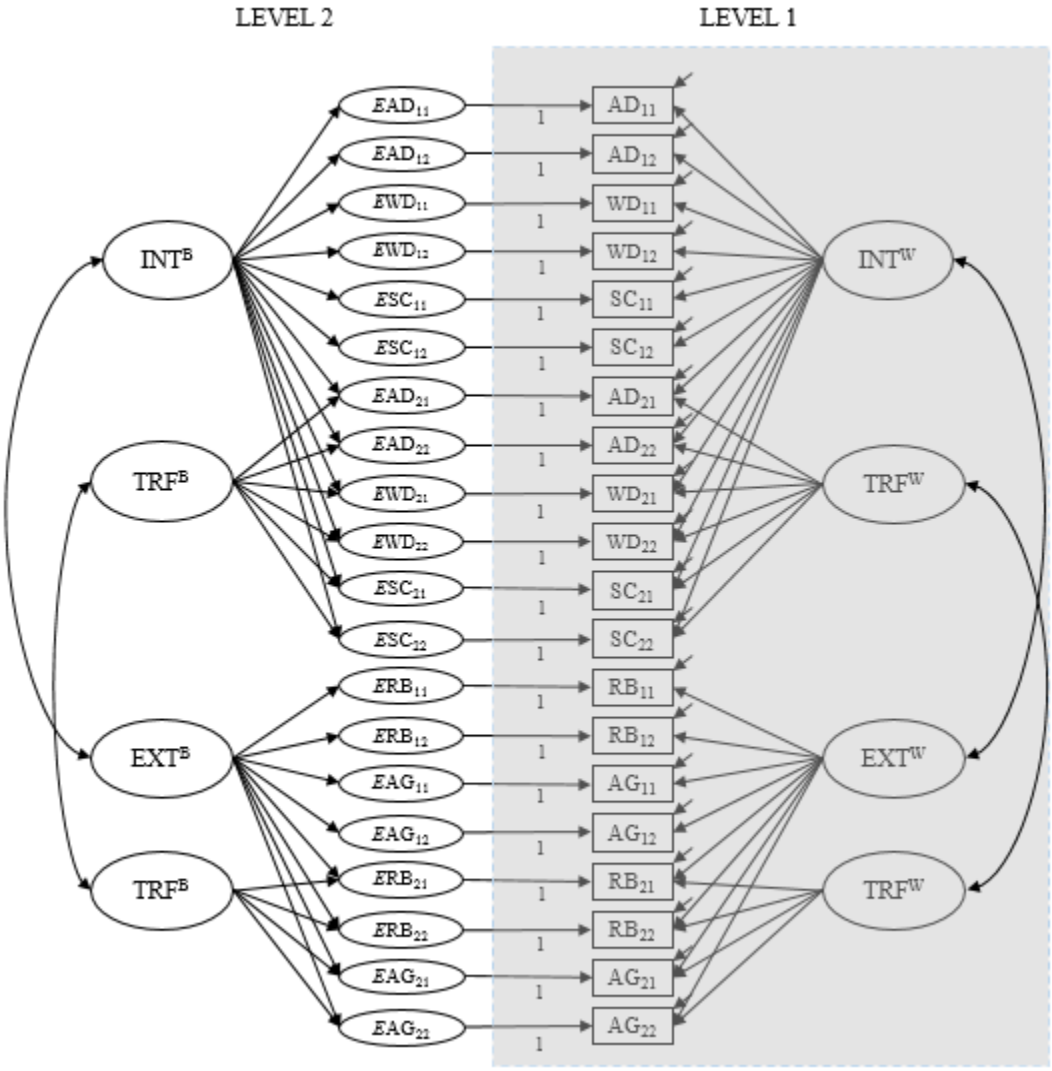

Figure 4. ML-CTC(M-1) model with two factors (Internalizing and Externalizing), two indicators and two methods (teachers – TRF and students) with YSR as the reference group. Within is the self-report or Level 1 and Between is the teacher report or Level 2. AD =Anxious/Depressed, WD = Withdrawn/Depressed, SC = Somatic Complaints, SP = Social Problems, TP = Thought Problems, AT = Attention Problems, RB = Rule Breaking Behaviors and, AG = Aggressive Behaviors.

References

Achenbach, T. M. (1991). *Manual for the child behavior checklist/4-18 and 1991 profile*. Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T. M. (2006). As others see us: Clinical and research implications of cross-informant correlations for psychopathology. *Current Directions in Psychological Science*, *15*(2), 94-98.

Achenbach, T. M., Becker, A., Döpfner, M., Heiervang, E., Roessner, V., Steinhausen, H. C., & Rothenberger, A. (2008). Multicultural assessment of child and adolescent psychopathology with ASEBA and SDQ instruments: research findings, applications, and future directions. *Journal of Child Psychology and Psychiatry, 49*, 251-275.

Achenbach, T. M., Dumenci, L., & Rescorla, L. A. (2002). Ten-year comparisons of problems and competencies for national samples of youth self, parent, and teacher reports. *Journal of emotional and Behavioral disorders*, *10*(4), 194-203.

Achenbach, T. M. & Edelbrock, C. (1983). *Manual for the child behavior checklist/4-18 and revised child behavior profile*. Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T. M., Krukowski, R. A., Dumenci, L., & Ivanova, M. Y. (2005). Assessment of adult psychopathology: Meta-analyses and implications of cross-informant correlations. *Psychological Bulletin*, *131*, 361-382.

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: implications of cross-informant correlations for situational specificity. *Psychological Bulletin*, *101*, 213-232.

Achenbach, T. M., & Rescorla, L. (2001). *Manual for the ASEBA school-age forms & profiles: An integrated system of multi-informant assessment.* Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families.

Achenbach, T. M., & Rescorla, L. A. (2007). *Multicultural supplement to the manual for the ASEBA school-age forms & profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families.

Albrecht, G., Veerman, J. W., Damen, H., & Kroes, G. (2001). The Child Behavior Checklist for group care workers: A study regarding the factor structure. *Journal of Abnormal Child Psychology, 29*(1), 83-89.

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

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: NY, The Guilford Press.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56,* 81-105.

Carretero-Dios, H., Eid, M., & Ruch, W. (2011). Analyzing multitrait-mulitmethod data with multilevel confirmatory factor analysis: An application to the validation of the State-Trait Cheerfulness Inventory. *Journal of Research in Personality, 45,* 153-164.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14,* 464-504.

Cole, D. (1987). Utility of confirmatory factor analysis in test validation research. *Journal of Consulting and Clinical Psychology, 55,* 584-594.

Coleman, M. C., & Webber, J. (2002). *Emotional and behavioral disorders: Theory and practice*. Boston, MA: Allyn & Bacon.

Cullinan, D., & Sabornie, E. J. (2004). Characteristics of emotional disturbance in middle and high school students. *Journal of Emotional and Behavioral Disorders, 12*, 157–167.

Dedrick, R. F., Greenbaum, P. E., Friedman, R. M., Wetherington, C. M., & Knoff, H. M. (1997). Testing the structure of the Child Behavior Checklist/4-18 using confirmatory factor analysis. *Educational and Psychological Measurement, 57*, 306-313.

Dedrick, R. F., Tan, T. X., & Marfo, K. (2008). Factor structure of the Child Behavior Checklist/6-18 in a sample of girls adopted from China. *Psychologicol Assessment, 20*(1), 70-75.

De Groot, A., Koot, H. M., & Verhulst, F. C. (1994). Cross-cultural generalizability of the Child Behavior Checklist cross-informant syndromes. *Psychological Assessment, 6*, 225-230.

Eid, M. (2000). A multitrait–multimethod model with minimal assumptions. *Psychometrika, 65,* 241–261.

Eid, M., Lischetzke, T., & Nussbeck, F. W. (2006). Structural equation model for multitrait-multimethod data. In M. Eid, & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 283-299). Washington, DC: American Psychological Association.

Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait–multimethod models: A multiple-indicator CTC(*M*–1) model. *Psychological Methods, 8*, 38–60.

Eid, M., Nussbeck, F. W., Geiser, C., Cole, D. A., Gollwitzer, M., & Lischetzke, T. (2008). Structural equation modeling of multitrait-multimethod data: Different models for different types of methods. *Psychological Methods, 13*, 230-253.

van der Ende, J., Verhulst, F. C., & Tiemeier, H. (2012). Agreement of informants on emotional and behavioral problems from childhood to adulthood. *Psychological Assessment, 24*, 293-300.

Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling: A Multidisciplinary Journal, 8*, 128-141.

Epanchin, B. (1991). Assessment of social and emotional problems. In J. L. Paul, & B. C. Epanchin (Eds). *Educating emotionally disturbed children and youth: Theories and practices for teachers* (pp. 307-349). New York: Macmillan Publishing Company.

Flamer, S. (1983). Assessment of the multitrait-multimethod matrix validity of likert scales via confirmatory factor analysis. *Multivariate Behavioral Research, 18,* 275-308.

Gomez, R., Vance, A., & Gomez, R. M. (2014). Analysis of the convergent and discriminant validity of the CBCL, TRF, and YSR in a clinic-referred sample. *Journal of Abnormal Child Psychology, 42,* 1413-1425.

Greenbaum, P. E., & Dedrick, R. F. (1998). Hierarchical confirmatory factor analysis of the Child Behavior Checklist/4–18. *Psychological Assessment, 10,* 149-155.

Gresham, F. M., Cook, C. R., Crews, S. D., & Kern, L. (2004). Social skills training for children and youth with emotional and behavioral disorders: Validity considerations and future directions. *Behavioral Disorders, 30*(1), 32-46.

Gresham, F. M., Elliott, S. N., Cook, C. R., Vance, M. J., & Kettler, R. (2010). Cross-informant agreement for ratings for social skill and problem behavior ratings: An investigation of the Social Skills Improvement System—Rating Scales. *Psychological Assessment, 22,* 157-166.

Gresham, F. M., & Kern, L. (2004). Internalizing behavior problems in children and adolescents. In R. Rutherford Jr., M. Quinn & S. Mathur (Eds.), *Handbook of research in emotional and behavioral disorders* (pp. 262-281). New York, NY: Guilford Press.

Grigorenko, E., Geiser, C., Slobodskaya, H., & Francis, D. (2010). Cross-informant symptoms from CBCL, TRF, and YSR: Trait and method variance in a normative sample of Russian youths. *Psychological Assessment, 22,* 893-911.

Hox, J., & Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling, 8,* 157-174.

Ivanova, M. Y., Achenbach, T. M., Dumenci, L., Rescorla, L. A., Almqvist, F., Weintraub, S., ... & Verhulst, F. C. (2007). Testing the 8-syndrome structure of the child behavior checklist in 30 societies. *Journal of Clinical Child and Adolescent Psychology, 36,* 405-417.

Ivanova, M. Y., Achenbach, T. M., Rescorla, L. A., Dumenci, L., Almqvist, F., Bathiche, M., ... & Verhulst, F. C. (2007). Testing the Teacher's Report Form syndromes in 20 societies. *School Psychology Review, 36,* 468-483.

Ivanova, M. Y., Achenbach, T. M., Rescorla, L. A., Dumenci, L., Almqvist, F., Bilenberg, N., ... & Verhulst, F. C. (2007). The generalizability of the Youth Self-Report syndrome structure in 23 societies. *Journal of Consulting and Clinical Psychology, 75,* 729-738.

Julian, M. W. (2001). The consequences of ignoring multilevel data structures in nonhierarichal covariance modeling. *Structural Equation Modeling,* 8, 325-352.

Kauffman, J. M. (2001). *Characteristics of emotional and behavioral disorders in children and youth* (7th ed.). Upper Saddle River, NJ: Merrill/Prentice Hall.

Kenny, D. A. (1976). An empirical application of confirmatory factor analysis to the multitrait–multimethod matrix. *Journal of Experimental Social Psychology, 12,* 247–252.

Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin, 112,* 165-172.

Koch, T., Schultze, M., Burrus, J., Roberts, R. D., & Eid, M. (2015). A multilevel CFA-MTMM model for nested structurally different methods. *Journal of Educational and Behavioral Statistics, 40,* 477-510.

Ladd, G. W., & Kochenderfer-Ladd, B. (2002). Identifying victims of peer aggression from early to middle childhood: analysis of cross-informant data for concordance, estimation of relational adjustment, prevalence of victimization, and characteristics of identified victims. *Psychological Assessment, 14*(1), 74-76.

Lambert, M. C., Knight, F., Taylor, R., & Achenbach, T. M. (1994). Epidemiology of behavioral and emotional problems among children of Jamaica and the United States: Parent reports for ages 6–11. *Journal of Abnormal Child Psychology, 22,* 113–128.

Lambert, M. C., Lyubansky, M., & Achenbach, T. M. (1998). Behavioral and emotional problems among adolescents of Jamaica and the United States: Parent, teacher, and self-reports for ages 12 to 18. *Journal of Emotional and Behavioral Disorders, 6,* 180–187.

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

Lane, K. L., Barton-Arwood, S. M., Nelson, J. R., & Wehby, J. (2008). Academic performance of students with emotional and behavioral disorders served in a self-contained setting. *Journal of Behavioral Education, 17*(1), 43-62.

Lane, K. L., Pierson, M., & Givner, C. C. (2004). Secondary teachers' views on social competence: Skills essential for success. *Journal of Special Education, 38,* 174–186.

Lee, S. W., Elliott, J., & Barbour, J. D. (1994). A comparison of cross-informant behavior ratings in school-based diagnosis. *Behavioral Disorders, 19*(2), 87-97.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural equation modeling, 9*(2), 151-173.

Maas, C., Lensvelt-Mulders, G., & Hox, J. (2009). A multilevel multitrait-multimethod analysis. *Methodology, 5*(3), 72-77.

Marsh, H. W. (1989). Confirmatory factor analysis of multitrait–multimethod data: Many problems and a few solutions. *Applied Psychological Measurement, 13,* 335–361.

Marsh, H. W. (1993). Multitrait-multimethod analysis: Inferring each trait/method combination with multiple indicators. *Applied Measurement in Education, 6,* 49-81.

Marsh, H. W., & Bailey, M. (1991). Confirmatory factor analyses of multitrait-multimethod data: A comparison of alternative models. *Applied Psychological Measurement, 15*(1), 47-70.

Marsh, H. W., & Byrne, B. M. (1993). Confirmatory factor analysis of multitrait–multimethod self-concept data: Between group and within-group invariance constraints. *Multivariate Behavioral Research, 28,* 313–349.

Marsh, H. W., & Grayson, D. (1995). Latent variable models of multitrait–multimethod data. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 177–198). Thousands Oaks, CA: Sage.

Marsh, H. W., & Hocevar, D. (1988). A new, more powerful approach to multitrait–multimethod analyses: Application of second-order confirmatory factor analysis. *Journal of Applied Psychology, 73,* 107–117.

McConaughy, S. H. (1993). Evaluating behavioral and emotional disorders with the CBCL, TRF, and YSR cross-informant scales. *Journal of Emotional & Behavioral Disorders, 1*, 40-52.

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

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

Ministry of Education, Youth and Sports (2012). Educational statistical digest of Belize 2011-2012. Belize City: Author.

Mooney, P., Ryan, J. B., Uhing, B. M., Reid, R., & Epstein, M. H. (2005). A review of self-management interventions targeting academic outcomes for students with emotional and behavioral disorders. *Journal of Behavioral Education, 14,* 203-221.

Moreland, A. D., & Dumas, J. E. (2008). Categorical and dimensional approaches to the measurement of disruptive behavior in the preschool years: A meta-analysis. *Clinical Psychology Review, 28,* 1059-1070.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Naslund-Hadley, E., Alonzo, H., & Martin, D. (2013). *Challenges and opportunities in the Belize education sector.* Inter-American Development Bank.

Nelson, J. R., Benner, G. J., Lane, K., & Smith, B. W. (2004). Academic achievement of K-12 students with emotional and behavioral disorders. *Exceptional Children, 71*(1), 59-73.

Polsgrove, L., & Smith, S. W. (2004). Informed practice in teaching self-control to children with emotional and behavioral disorders. In R. Rutherford Jr., M. Quinn & S. Mathur (Eds.), *Handbook of research in emotional and behavioral disorders* (pp. 399-425). New York, NY: Guilford Press.

Reef, J., van Meurs, I., Verhulst, F. C., & van der Ende, J. (2010). Children's Problems Predict Adults' DSM-IV Disorders Across 24 Years. *Journal of the American Academy of Child & Adolescent Psychiatry*, *49*, 1117-1124.

Renk, K. (2005). Cross-informant ratings of the behavior of children and adolescents: The "gold standard". *Journal of Child and Family Studies*, *14*, 457-468.

Renk, K., & Phares, V. (2004). Cross-informant ratings of social competence in children and adolescents. *Clinical Psychology Review*, *24*, 239-254.

Reynolds, C., & Kamphaus, R. (2004). *Behavioral assessment system for children* (2nd ed.). Minneapolis, MN: Pearson Assessment.

Richardson, G. A., & Day, N. L. (2000). Epidemiologic considerations. In M. Hersen & R. T. Ammerman (Eds.), *Advanced abnormal child psychology* (2nd Ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Sparrow, S., Cichetti, D., & Balla, D. (2005). *Vineland adaptive behavior scales* (2nd ed.) Minneapolis, MN: Pearson Assessment.

Synhorst, L. L., Buckley, J. A., Reid, R., Epstein, M. H., & Ryser, G. (2005). Cross informant agreement of the Behavioral and Emotional Rating Scale-(BERS-2) parent and youth rating scales. *Child & Family Behavior Therapy*,*27*(3), 1-11.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). New York, NY: Allyn and Bacon.

Taylor, T. K., & Biglan, A. (1998). Behavioral family interventions for improving child-rearing: A review for clinicians and policy makers. *Clinical Child and Family Psychology Review, 1*(1), 41–60.

United Nations, Departement of Economic and Social Affairs, Population Division. (2013). *World population prospects: The 2012 revision, key findings and advance tables.* Working Paper No. ESA/P/WP.227. Retrieved Feb 2, 2014, from http://esa.un.org/wpp/Documentation/pdf/WPP2012_%20KEY%20FINDINGS.pdf

Webster-Stratton, C., Reid, M. J., & Hammond, M. (2004). Treating children with early-onset conduct problems: Intervention outcomes for parent, child, and teacher training. *Journal of Clinical Child and Adolescent Psychology, 33*, 105-124.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9*(1), 1-26.

Witt, J. C., Vanderheyden, A. M., & Gilbertson, D. (2004). Instruction and classroom management: Prevention and intervention research. In R. Rutherford Jr., M. Quinn & S. Mathur (Eds.), *Handbook of research in emotional and behavioral disorders* (pp. 426-445). New York, NY: Guilford Press.

Wothke, W. (1995). Covariance components analysis of the multitrait–multimethod matrix. In P. E. Shrout (Ed.), *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske* (pp. 125–144). Hillsdale, NJ: Erlbaum.

Yu, C. Y., & Muthen, B. (2002, April). *Evaluation of model fit indices for latent variable models with categorical and continuous outcomes*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

APPENDIX A

EXTENDED LITERATURE REVIEW

Behavioral and emotional problems (BEPs) have been viewed as a difficult construct to define. BEPs, as a concept, are influenced by several factors: variation in individuals' tolerance ranges for behavior, theoretical model differences, terminology associated with emotional problems, and the sociological parameter of the problem (Coleman & Webber, 2002). The abovementioned factors contribute to the difficulty in defining BEP. A public law definition by the U. S. Department of Education (Federal Register, 1981) was adopted from Eli Bower's 1957 definition. The proposed definition (of what was then called emotional disturbance and now referred to as BEPs) has been adopted by many state educational departments and read as:

> Seriously emotionally disturbed is defined as follows: (i) the term means a condition exhibiting one or more of the following characteristics over a long period of time and to a marked degree, which adversely affects educational performance: (a) an inability to learn which cannot be explained by intellect, sensory, or health factors; (b) an inability to build or maintain satisfactory interpersonal relationships with peers and teachers; (c) inappropriate types of behavior or feelings under normal circumstances; (d) a general pervasive mood of unhappiness or depression; (e) a tendency to develop symptoms or fears associated with personal or school problems. (ii) The term includes children who are schizophrenic. The term does not include children who are socially maladjusted, unless it is determined that they are seriously emotionally disturbed. (Federal Register, 1981 as cited in Coleman & Webber, 2002, p. 27)

Bower (1982) later voiced concerns with the federal definition of BEP. In his original definition, Bower did not include 'seriously' in reference to emotionally disturbed children. The proposed definition excluded mildly and moderately emotionally disturbed children. Bower was also

concerned with the federal definition excluding children who are socially maladjusted. To that end the National Mental Health and Special Education Coalition formed working groups to propose an alternate definition. Many professionals and advocacy groups endorsed the proposed definition (Forness & Knitzer, 1992). Currently, Individuals with Disabilities Education Improvement Act (IDEA, 2002)) adopted a revised version of the proposed definition by the working group. BEP under the IDEA is as follows:

> A condition exhibiting one or more of the following characteristics over a long period of time and to a marked degree that adversely affects a child's educational performance:
>
> (A) An inability to learn that cannot be explained by intellectual, sensory, or health factors
>
> (B) An inability to build or maintain satisfactory interpersonal relationships with peers and teachers
>
> (C) Inappropriate types of behavior or feelings under normal circumstances
>
> (D) A general pervasive mood of unhappiness or depression
>
> (E) A tendency to develop physical symptoms or fears associated with personal or school factors. (IDEA, 2002)

Kauffman (2001) argues that an unsettled definition compounds the difficulty to accurately and reliably identify students with BEP. Early reports from the National Research Council (2002) and the U. S. Department of Education (2005) suggest that approximately one percent of public school students in the United States receive special education. Results from other studies (Costello, Egger, & Angold, 2005; Kauffman & Landrum, 2009) argue that the prevalence of BEP is at least five times greater than suggested by earlier studies (Kauffman, 2009). The *31st Annual Report for Congress* stated that in 2007, BEP was the fourth most prevalent disability

category for children and students age six through 21 years old (U. S. Department of Education, 2012). The result of having a widely acceptable definition allows for proper categorization and ultimately creating a path to assist children who are affected by BEP.

While reviewing the BEP literature, it is clear the concept of BEP in children (hereafter includes adolescents) has evolved because of the lack of clarity in its definition. For the purpose of this review BEP includes literature from emotional disturbance, emotional and behavioral disorders and behavioral problems. The previously stated concepts are similar in nature but may differ in terms of chronicity, frequency and severity. The concepts are used in the literature based on the whether the author is a practitioner in school, a clinician or a researcher in the field of emotional and behavioral children. Irrespective of the name used to describe the emotions or behaviors of children, the measurement tools used are the same.

Behavior and Emotional Problems and the School

Behavioral and emotional problems (BEPs) are known to affect children's ability to shape and maintain effective social relationships with peers, teachers, parents and family (Cullinan & Sabornie 2004; Gresham, Cook, Crews, & Kern, 2004; Landrum, Tankersley, & Kauffman, 2003; Reid, Gonzalez, Nordness, Trout, & Epstein, 2004; Walker, Irvin, Noell, & Singer, 1992). Negative behaviors are often disruptive and lead to alienation from peers and denied learning opportunities from adults (Kauffman, 2001). Ultimately, unless BEP behaviors are curtailed, children's functionality in society will be impaired.

Children with BEPs are susceptible to negative educational and social outcomes at school. Nelson, Benner, Lane, and Smith (2004) examined K-12 public school students and found students with emotional and behavioral disorder had large academic achievement deficits. Students showed severe shortfalls in reading, written expression, and mathematics achievement

even while controlling for age of onset of the problems (Nelson et al., 2004). Lane, Barton-Arwood, Nelson, and Wehby (2008) found similar results for both elementary and secondary students with behavior problems. BEP students typically score below the 25th percentile on reading, mathematics and written expression outcomes (Lane et al., 2008).

Interpersonal relationships, similar to academic outcomes, can be difficult for children with BEPs (Kauffman, 2001). BEPs in children tend to be effective predictors of problem behaviors in adulthood (Reef, van Meurs, Verhulst & van der Ende, 2010). In a 24-year longitudinal study, Reef et al. (2010) found that children who were reported (by parents) to have deviant behaviors tended to exhibit disruptive behaviors as adults; children with conduct disorder had mood and disruptive disorder at adulthood and; children with anxiety showed further anxiety problems in middle adulthood. Moreland and Dumas (2008) similarly argued that behavior problems beginning at preschool age tends to be associated with life-long challenges. Gresham, et al. (2004) found social skills training effective in combatting several behavioral difficulties in children and youth. Early social skills interventions can prevent adult delinquency and substance abuse (Taylor & Biglan, 1998), decreased negative behavior, and increased prosocial skills with peers (Webster-Stratton, Reid, & Hammond, 2004).

However, with self-management interventions (self-monitoring, self-evaluation, strategy instruction and goal setting), students with BEPs can produce positive academic outcomes (Mooney, Ryan, Uhing, Reid, & Epstein, 2005). Students with emotional and behavioral problems show marked improvements in academic skills when introduced to these interventions. Similarly, teacher instruction and classroom management have shown to improve students' academic outcomes (Witt, Vanderheyden, & Gilbertson, 2004). Witt et al. argued that a deeper focus on instructional design (for example feedback and sequencing) would result in teachers

having more success redirecting students' academic and behavioral outcomes. Teachers have identified that the school process can also be successfully navigated if students with behavior problems enhance self-control or cooperation skills (Lane, Pierson, & Givner, 2004; Lane, Wehby, & Cooley, 2006; Polsgrove & Smith, 2004).

Understanding the Characteristics of Behavioral and Emotional Problems

Behavioral and emotional problems (BEPs) are characterized into two factors that conceptualize deviant behaviors (Coleman & Webber, 2002). The two factors are labeled internalizing and externalizing behaviors. Internalizing behaviors represent introverted problems, directed inwardly to the individual. Children exhibiting internalizing problems tend to present subtle markers and typically go unnoticed (Gresham & Kern, 2004). Examples of internalizing behaviors include worries, social withdrawal, depression, anxiety, and obsessive-compulsive behaviors. On the other hand, externalizing behavior patterns represent behaviors that are directed outwardly. Children with externalizing BEPs are likely to be more impulsive while internalizing children are more reflective. Examples of externalizing behaviors are disobedience, aggression, disruption, impulsivity, and temper tantrums.

There are different theoretical approaches used to explain internalizing and externalizing behaviors in children. Behavior checklists, tools used to assess BEPs, are rooted in behavioral psychology (Epanchin, 1991). The behavioral model will be reviewed in explaining BEPs in children.

Behavioral model

The general orientation of the behavioral model in addressing BEPs follows that **c**hildren exhibit behaviors that are judged as inappropriate because the behaviors occur frequently or insufficiently (Cullinan, 2002). Further tenets of the behavioral model suggest that children learn

behaviors through conditioning (operant and respondent) and observational learning (social learning) – where consequences of the behaviors determine the frequency of the repetition of the behavior. Behaviorists believe there are few internal unobservable behaviors (Coleman & Webber, 2002). Additionally, normal and abnormal behaviors are learned and maintained in the same way (Schroeder & Riddle, 1991).

Ivan Pavlov and the salivating dog, in 1902, popularized the behavioral framework of respondent or classical conditioning. In an experiment, Pavlov realized that pairing a neutral stimulus (the sound of a bell) and an unconditioned stimulus (meat powder) can elicit a behavior – conditioned response (salivation) in his dog. Pavlov presented the neutral stimulus first, followed by the unconditioned stimulus. After several pairing of the neutral stimulus and unconditional stimulus, the dog would begin salivating once the neutral stimulus was presented. Early behaviorists such as Pavlov, argue children learn new behaviors through the pairing of neutral and unconditioned stimuli thus eliciting a reflex response (Coleman and Webber, 2002).

Behavior modification practitioners today, follow the operant conditioning frame of the behavioral model. They assume the principle of reinforcement drives all behavior (Bauer & Shea, 1999). Operant conditioning was made popular by E. L. Thorndike and B. F. Skinner. Initially known as instrumental conditioning, operant conditioning is centered on the control of reinforcement or punishment. Operant conditioning is based on the law of effect – behaviors are likely to increase or continued when followed by a rewarding consequence, and behaviors are likely to decrease when the consequences are unrewarding or punishable. Under this framework BEPs in children are explained in terms of rewards and punishments.

From a social learning perspective, normal and maladaptive behaviors in children are shaped (and reinforced) by social context (Bandura, 1986). Bandura proposed reciprocal

determinism that can be used to explain psychopathology. He argued that the person, behavior and environment work together interactively. For example, aggression - an externalizing behavior, results from aggressive behaviors occurring in the children's social interactions with siblings, parents, peer and teachers. Children learn the aggressive behaviors simply by observing (Furlong, Morrison, & Jimerson, 2004). Once a behavior has been observed, children are likely to be affected in one of three possible ways: novel responses to the behavior are received, behavior becomes repressed or disinhibited, or previously learned responses may be reinforced (Coleman & Webber, 2002; Schroeder & Riddle, 1991). It is important to note that learned behavior may not be performed unless there is an associated consequence or incentive.

Assessment Methods: Classification of BEPs

Research conducted in abnormal behaviors has been largely diagnostic (Bird, 1996). Diagnostic assessment classifies children categorically, as disordered or not disordered. In categorical classification, BEPs are "maladaptive and distressing behaviors, emotions, and thoughts that is qualitatively different from normality" (Cullinan, 2004, p. 33). Empirically based assessment (also known as dimensional) represent children's behavior on a continuous dimension (Achenbach et al., 2008). Behaviors, emotions and thoughts are experienced by all people but on a continuum (Cullinan, 2004).

Categorically-based Methods

Categorical methods of assessment for BEPs are used to screen or identify people most likely needing a diagnosis and predominantly used for clinical interventions. Clinicians and researchers use categorical methods when attempting to make an initial determination of the presence or absence of a behavior or disorder. Arguably the most common example of the categorical approach to disorders is the *Diagnostic and Statistical Manual of Mental Disorders*

(*DSM-V*; American Psychiatric Association, 2013). The DSM was created for psychiatrists to classify health abnormalities, diseases or nosology (American Psychiatric Association, 1952). A health abnormality is linked to symptoms and signs. Symptoms are experienced and subjectively reported by the individual. Signs are changes observed in body functioning. The reoccurrence of these symptoms and signs are known as syndromes (Cullinan, 2004).

The focus of the DSM, in its earlier versions, was on distinguishing between adult disorders such as schizophrenia and manic-depression. It was not until the *DSM-III* (American Psychiatric Association, 1980) that new diagnostic categories were included to assess childhood disorders (Achenbach & McConaughy, 1996).

Empirically-based Methods

Brown and Barlow (2009) highlighted that categorically-based assessments produce favorable reliability estimates, generally. The authors, however, argued the categorically-based method shows relatively poor ability to distinguish emotional disorder. The use of the *DSM*'s categorical approach to classifying diagnoses possesses measurement errors when applying a categorical cutoff and; distorts diagnostic rates owing to overlapping criteria of the disorders (Brown & Barlow, 2009).

Brown and Barlow (2009) proposed a move from the categorically-based method of assessing anxiety and mood disorder to using an empirically-based method. The *Achenbach System of Empirically Based Assessment (ASEBA;* Achenbach, 1998; Achenbach & Rescorla, 2001) is the most popular empirically-based method for assessing BEPs in children (Merrell, 2008). The empirically-based method of assessing BEPs allows for the detection of chronicity, frequency and severity. Another advantage of the empirically-based method is having

standardized assessments for cross-group comparisons. Assessments, for example the *ASEBA* instruments, allow for measuring multiple behaviors in one setting (Cullinan, 2004).

Measuring Behavioral and Emotional Problems in Children

The Child Behavior Checklist (CBCL) is a measure developed by *ASEBA*. The CBCL school-age forms were first created as the CBCL/4-18 (Achenbach, 1991; Achenbach & Edelbrock, 1983). The CBCL assesses the degree of informants' consistency to report behavioral and emotional problems in children/adolescents (Achenbach, McConaughy, & Howell, 1987). Since the inception of the CBCL/4-18, a number of validation studies of the measures have been conducted (Albrecht, Veerman, Damen, & Kroes, 2001; Greenbaum & Dedrick, 1998). The CBCL/4-18 was revised and is now known as CBCL/6-18 with a change in the age range (Achenbach & Rescorla, 2001).

The CBCL/4-18 and the revised CBCL/6-18 has a similar structure, consisting of eight syndrome scales. The eight syndrome scales are: Withdrawn, Somatic Complaints, Anxious/Depressed, Social Problems, Thought Problems, Attention Problems, Delinquent/Rule-Breaking Behavior, and Aggressive Behavior (Achenbach, 1991; Dedrick, Greenbaum, Friedman, Wetherington, & Knoff, 1997; Dedrick, Tan, & Marfo, 2008; De Groot, Koot, & Verhulst, 1994; Ivanova, Achenbach, Dumenci, et al., 2007). The eight syndrome scales were shown to have a higher-order factor structure with Delinquent/Rule-Breaking and Aggressive scales as indicators of *Externalizing Disorders* and Withdrawn, Somatic, and Anxious–Depressed scales as indicators of *Internalizing Disorders* (Achenbach, 1991; Greenbaum & Dedrick, 1998). The names of the grouping are due to within-self problems (internalizing) and for externalizing problems concerning conflicts with others and with participants' expectations (Achenbach & Rescorla, 2001).

The CBCL school-age forms are comprised of three components/forms: CBCL/6-18, completed by parents or surrogates, Youth Self-Report (YSR), and Teacher's Report Form (TRF). The CBCL is a multicultural problems assessment instrument with normative samples for over 40 societies (Achenbach & Rescorla, 2007). The syndromes are named to represent the description of the problems and not as a diagnosis (Achenbach, 1991; McConaughy, 1993). However, it is suggested that the DSM-IV diagnoses of similar descriptions are highly correlated with the syndromes (Achenbach & Rescorla, 2001).

CBCL (Parent/Guardian Form)

The current version of the CBCL has been normed for children ages 11-18 years old. Parents or guardians of children complete the CBCL by describing the children's behaviors that are observed. The first two pages request parents describe activities in which their children participate, as well as the frequency and the effectiveness of participating in those activities compared to children of the same age. The final two pages contain 118 statements describing the child. Items are measured on a 3-point scale; "Not True", "Somewhat or Sometimes True" and "Very True or Often True". The number of items associated with each CBCL (Parent/Guardian Form) syndromes are: Withdrawn (8 items), Somatic Complaints (11 items), Anxious/Depressed (13 items), Social Problems (11 items), Thought Problems (15 items), Attention Problems (10 items), Delinquent/Rule-Breaking Behavior (17 items), and Aggressive Behavior (18 items). Parents are required to rate the child within the last six months (e.g. "My child argues a lot").

Youth Self-Report

The current form of the YSR has been normed for children ages 11-18 years old (Appendix A). Youths complete the YSR describing their functioning. Demographic information and extra-curricular activities are requested on the first two pages. The next two pages contains

118 statements describing the youth. Items are measured on a 3-point scale; "Not True", "Somewhat or Sometimes True" and "Very True or Often True". The items were coded so that higher scores reflected more of the construct. The number of items associated with each YSR syndromes are: Withdrawn (8 items), Somatic Complaints (10 items), Anxious/Depressed (13 items), Social Problems (11 items), Thought Problems (12 items), Attention Problems (9 items), Delinquent/Rule-Breaking Behavior (15 items), and Aggressive Behavior (17 items). Items are written in the first person (e.g. "I enjoy a good joke").

Teacher Report Form

The current revised edition of the TRF has been normed for children ages 6 -18 years old. The TRF is completed by teachers or other school personnel (e.g. assistant teacher, principal, administrator or counselor) describing children's functioning at school. Like the YSR, the teacher form consists of 118 statements describing the youth. Similarly, the items were also measured on the "Not True", "Somewhat or Sometimes True" and "Very True or Often True" 3-point scale. Teachers are required to rate how well the behavioral, emotional and social problem items apply to youth (e.g. "Has difficulty learning"). Higher scores reflect more of the construct. The number of items associated with each YSR syndromes are: Withdrawn (8 items), Somatic Complaints (9 items), Anxious/Depressed (16 items), Social Problems (11 items), Thought Problems (10 items), Attention Problems (26 items), Delinquent/Rule-Breaking Behavior (12 items), and Aggressive Behavior (20 items).

Reports by Multiple Informants/Cross-informants

Collecting information from children across different situations and environments, ultimately using different informants, offers a more complete picture of children's behaviors. Behaviors should be reported consistently across situations and environments and behaviors

should least moderately correlate across informants (Achenbach, et al., 1987; Merrell, 2000; Synhorst, Buckley, Reid, Epstein, & Ryser, 2005). Multiple methods/sources reporting behaviors provides "valuable information in the examination of…emotional and behavioral functioning in children and adolescents" (Renk & Phares, 2004, p. 240). Richardson and Day (2000) and Renk (2005) argued for the use of multiple informants along with self-reports to measure behaviors. The authors suggest children's behaviors tend to be specific to their situations and collecting data from multiple environments contributes to a holistic understanding of behavior. Additionally, raters may observe different types and severity of behaviors that otherwise would have not been self-reported.

Low correlations among multiple informants have been found when using the CBCL for: psychopathology (Reynolds & Kamphaus, 2004), adaptive behavior (Sparrow, Cichetti, & Balla, 2005), behavioral problems (Lee, Elliott, & Barbour, 1994) and social skills (Gresham, Elliott, Cook, Vance, & Kettler, 2010). However, Ladd & Kochenderfer-Ladd (2002), similar to Achenbach et al., 1987, found moderate agreement between youth report of behavior and adult reports. A meta-analysis conducted by Achenbach, Krukowski, Dumenci, and Ivanova (2005) found moderate to high correlations between self-report and informants on parallel instruments for substance abuse, internalizing and externalizing problems. Results from the meta-analysis revealed high rater agreement when externalizing problems are being measured. Additionally, the mean cross-informant correlations were not significantly different for externalizing and internalizing problems. Similarly, Achenbach (2006) found moderate cross-informant agreement for adults with aggression and rule-breaking problems and for adults with anxiety, depression and withdrawal problems.

APPENDIX B

EXTENDED METHODOLOGICAL OVERVIEW

Psychological concepts are difficult to measure explicitly; therefore, items are combined to create measures of the underlying constructs. This measurement difficulty drives researchers to develop methods where instruments are validated. The Standards for Educational and Psychological Testing (American Educational Research Association, 2014) highlighted this validation process as the extent to which a measure is supported by evidence and theory (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Messick (1989) defined validity as "the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores" (p. 13).

Messick's (1989) conceptualization of validity has had a number of objections (Borsboom, Mellenbergh, & Heerden, 2004; Lissitz & Samuelson, 2007) and modifications (Hubley & Zumbo, 2011; Shepard, 1993). In this unified understanding of validity, Messick (1995) also highlighted six aspects of construct validity: (a) evidence of content relevance; (b) theoretical rationales for the observed consistencies; (c) fidelity of the scoring structure; (d) extent to which score properties and interpretations are generalizable across groups and settings; (e) convergent and discriminant evidence from multitrait-multimethod (MTMM) comparisons; and (f) value implications of score interpretation (p. 745). Establishing (a), (b) and (f) requires a deep theoretical agreement across members in the field. Psychometric evaluations are required to establish evidence of (c), (d) and (e). These aspects of validity address the magnitude of similarity of the item responses to the theoretical underpinning of the constructs.

The term construct validity was used by Cronbach and Meehl (1955) in addressing the debate of validation of psychological tests. However, it was Campbell and Fiske (1959) who

made the most significant contribution to the popularity of the concept. MTMM analyses are arguably the most widely used methodological analysis of convergent and discriminant validity of psychological measures. The MTMM matrix combines a cannon of traits with a cannon of measurement methods (Kenny & Kashy, 1992).

MTMM matrices were first analyzed by Campbell and Fiske (1959) by eyeballing the correlation matrix following three rules. First, a measure is said to have convergent validity when measures of the same trait or ability have high correlations. Second, methods of collecting the data should discriminate different traits. This is evidence of discriminant validity. These correlation coefficients should be low, and lower in relation to the same-trait, different methods correlation (Kenny & Kashy, 1992). Finally, in order to assess method variability the correlation between same-method and different trait should be similar to the different-method, different-trait correlations.

Multimethod examinations are preferred to single method evaluation of programs (Eid, 2006; Eid, Nussbeck, Geiser, Cole, Gollwitzer, & Lischetzke, 2008). The popularity of multimethod research opened an avenue for a number of proposed methodological approaches for analyzing multimethod research. A number of methodologies exist for analyzing MTMM data. There are three general categories covering the methodologies: correlation and association methods, analysis of variance (ANOVA) models and, latent variables models (Eid, 2006). Correlation and association models directly stem from the Campbell and Fiske (1959) correlation matrix idea. The *direct product model* (Swain, 1975) was developed under the umbrella of the correlation and association models. In the direct product model, the correlations represent the relationship between two traits, assessing discriminant validity while convergent validity is assessed as the relationship between two methods. Browne (1984) expanded Swain's work to the

composite direct product model; where considers the influence of measurement errors. Although the correlation and association models are useful they lack the ability to partition variances into trait and method portions.

The second of models used to analyze MTMM data are the ANOVA models. Under the ANOVA designs, convergent validity is determined by evaluating the variance component and intraclass correlation coefficient (Tinsley & Weiss, 2000). ANOVA models have since been extended with generalizability theory focusing on different measurement conditions or facets. Following the linear nature of ANOVA, multilevel analysis, is also used to analyze MTMM data (Hox, 2010). However, ANOVA designs do not detect trait specific method influences (Eid, 2004).

Messick (1989) argued for the use of confirmatory factor analysis (CFA) in assessing construct validation. CFA, more generally latent variable or structural equation modeling (SEM), when being used to analyze MTMM data, separates measurement errors from individual differences due to method and trait effects. CFA MTMM models allow for testing relationships of trait and method variables with other latent variables. Additionally, all models can empirically test the underline assumptions of the models. Subsequently, CFA models, have been mostly applied to analyzing MTMM data (Carretero-Dios, Eid, & Ruch, 2011; Cole, 1987; Eid et al., 2008; Flamer, 1983; Kenny & Kashy, 1992; Marsh & Bailey, 1991; Widaman, 1985).

Confirmatory Factor Analysis MTMM

Confirmatory factor analysis (CFA) is associated with testing theory (Messick, 1989). In the context of the present study, CFA assesses how well data fit a hypothesized measurement model (Tabachnick & Fidell, 2007). In a measurement model, specific items are loaded/weighted on theorized factors. This item-factor loading is referred to as the factor structure. In CFA-

MTMM models the underlying factor structure is decomposed. Individual differences are partitioned with the identification of the influence of the measured trait, method and the error (or unique) components. CFA-MTMM is predicated on the same assumptions outlined by Campbell and Fiske (1959) that "each measure loads on only one trait and one method factor, and that the covariances between trait and method factors are zero" (Maas, Lensvelt-Mulders, & Hox, 2009, p. 73).

There have been a wide range of CFA-MTMM models developed in the last three decades. Widaman (1985) posited a taxonomy of 16 CFA-MTMM models comparing four traits and four methods. Since then, other models have been developed (Eid, 2000; Eid, Lischetzke, Nussbeck, & Trierweiler, 2003; Kenny & Kashy, 1992; Marsh & Byrne, 1993; Marsh & Grayson, 1995; Marsh & Hocevar, 1988; Wothke, 1995). The correlated trait-correlated method (CT-CM; Marsh, 1989; Marsh & Grayson, 1995) and the correlated trait-correlated uniqueness model (CT-CU; Kenny, 1976) are the most used models (Eid et al., 2008).

In the CT-CM model, an observed variable designed to measure a trait is decomposed into a trait component, a method component and a residual component, not explained by both trait and method factors (Figure 1a). As per the name of the model, the within trait and within methods are allowed to correlate, but CT-CM model assumes the trait and method factors are uncorrelated. The trait and method factors each, independently, contribute to the explanation of the variance in the observed variable. The CT-CM model is known to have estimation problems since these models yield negative variances. Additionally, interpretation problems exist whenever methods factors are correlated (Eid et al., 2003; Marsh, 1989). When using the CT-CM model there is uncertainty of the conditions under which the trait and method factors are assumed uncorrelated. The CT-CM model also assumes the effect of one method is generalizable

across all uses of the method. This is a seemingly restrictive assumption, particularly with psychological measures (Eid et al., 2003).

The CT-CU was developed to address the aforementioned problems of the CT-CM model, namely the estimation and identification problems and the restrictive assumption. In the CT-CU model, there is no methods factor, however, the measures sharing common methods have covarying errors (Figure 1b). Residuals from the same methods may covary, therefore eliminating the assumption (made by CT-CM models) that method effects from one method are homogenous across all methods. On the other hand, between methods covariation is not specified, which is highly likely in research using multiple peer ratings. Furthermore, there are additional limitations when applying CT-CU models since there tends to be an underestimation of the proportion of variance of the observed variable being explained by method factors that is not due to measurement error. Finally, the relationship between error-free methods effects and external variables cannot be examined because of the absence of unspecified methods effects (Eid, et al., 2003).

An alternate model, correlated trait-correlated method minus one model [CT-C(*M*-1)] (Eid, 2000; Eid et al., 2003) was proposed, addressing shortcomings of the previous models. The CT-C(*M*-1) model is a unique variant of the CT-CM model. In the CT-C(*M*-1) modeling approach, one method is selected as a reference method and contrasted with all other methods (Figure 1c). This omitted method is also known as the comparison standard (Brown, 2006). The self-report method is often the most frequently used reference/comparison method. The main premise of the CT-C(*M*-1) model follows the notion that for each trait measured, the true score variables of the reference method are regressors in a latent regression where the dependent measures are the true score variables of the non-reference method(s) (Brown, 2006; Eid et al.,

2003). The method effects then become residuals common to all measured variables by the same method.

The methods used to collect data should be the determining factor for selecting models to analyze MTMM data (Eid et al., 2008). One criterion in determining the type of model is whether the method is interchangeable or structurally different. Eid et al. (2008) outlined the differences between interchangeable and structurally different raters. Interchangeable methods (raters) involve randomly selecting participants to rate an individual. For example, several students were randomly selected from a class and asked to rate the teacher's performance. Alternatively, if a teacher completes a self-report form and a peer and supervisor also rate the teacher then the raters are deemed to be structurally different. The ASEBA forms are considered structurally different because the teachers and parents both rate the behaviors of the child, who also rate themselves. When comparing structurally different methods the CT-C(*M*-1) is the most appropriate model (Eid, Lischetzke & Nussbeck, 2006; Koch, Schultze, Burrus, Roberts, & Eid, 2015).

Model Fit

Model fit indices are used in CFA to determine the magnitude of the hypothesized factor structure and data fit relationship. More technically, fit denotes the model's capability of reproducing the variance-covariance matrix (Kenny, 2014). Numerous fit indices are used in CFA. The Chi-Square test (χ^2) is a classic goodness-of-fit index. When models possess a small sample size ($n < 200$) is reasonable to use the χ^2 as a measure of fit (Kenny, 2014). A statistically significant χ^2 (sig. < .05) supports the alternative hypothesis that the estimates in the model do not reproduce the variances and covariances in the data (Brown, 2006) indicating the model does not fit the data well. Non-significant χ^2 indices are preferred.

With large samples, as is often the case in multivariate analyses, χ^2 will almost always be statistically significant (Barrett, 2007). The chi-square statistic is therefore seldom used in research as the sole statistic in assessing model fit. In addition to the χ^2 other model fit indices are used that can be characterized in three broad categories: (a) absolute fit indices that are used to evaluate the proportions of the sample data covariances being explained by the model, where the best fit is zero; (b) comparative or incremental fit indices that are used to evaluate improvement in fit of the sample data compared with a baseline model and; (c) parsimony fit indices that include a correction to the model degrees of freedom (df_M) for the complexity of the model (Brown, 2006; Kenny, 2014; Kline, 2011).

The absolute fit index in the present study is the standardized root mean square residual (SRMR). The SRMR can be viewed as the average error or difference between the sample data variance-covariance matrix and the model implied variance-covariance matrix. SRMR is the difference between observed and predicted correlations (Kenny, 2014). Perfect fit is indicated by SRMR = 0, and higher values indicate poor fit. Like χ^2, SRMR is an indicator of how poorly a model fits the data. SRMR is calculated by summing the squared elements of the residual correlation matrix, dividing by the total elements in the matrix, and taking the square root of the result (Brown, 2006) and can be expressed as follows (1):

$$SRMR = \sqrt{\frac{\sum_{1}^{u}(model-based\ standardized\ residual)^2}{u}} \qquad (1)$$

where $u = p(p+1)/2$ is the number of unique variances/covariances among the p variables in the model. Hu and Bentler (1999) suggest SRMR values less than .08 may be considered good fit to the data.

The comparative fit index (CFI; Bentler, 1990) and the Tucker-Lewis Index (TLI; Tucker & Lewis, 1973) are two popular comparative fit indices. These indices compare the target model to a baseline model. The baseline model has covariances of the indicators set to zero with no constraints on the variances (Brown, 2006). The formula for CFI is given in (2):

$$CFI = 1 - \frac{\max\left[\left(\chi_T^2 - df_T\right), 0\right]}{\max\left[\left(\chi_B^2 - df_B\right), \left(\chi_T^2 - df_T\right), 0\right]} \quad (2)$$

where max indicates the use of the largest value (the difference between the χ^2 and *df* or 0), χ_T^2 is the χ^2 value of the target model being evaluated, df_T is the *df* of the target model, χ_B^2 is the χ^2 value for the base or null model, df_B is the *df* of the null model. The CFI value is normed and ranges between 0.0 and 1.0. TLI is given in (3):

$$TLI = \frac{\left[\left(\chi_B^2 / df_B\right) - \left(\chi_T^2 / df_T\right)\right]}{\left[\left(\chi_B^2 / df_B\right) - 1\right]} \quad (3)$$

where, similar to CFI, χ_T^2 is the χ^2 value of the target model being evaluated, df_T is the *df* of the target model, χ_B^2 is the χ^2 value for the base or null model, df_B is the *df* of the null model. Although interpreted similarly to CFI, the TLI value is non-normed and therefore does not range from 0.0 to 1.0. Values greater than or equal to .95 constitute good fit while .90 represents marginal fit (Hu & Bentler, 1999).

Root Mean Square Error of Approximation (RMSEA; Steiger, 1990) is a commonly used parsimony correction index assessing the degree to which a model fits well in the population (Brown, 2006). RMSEA is highly related to chi-square and should be reported along with other indices (Fan & Sivo, 2007; Kenny & McCoach, 2009). The computation of RMSEA is given in (4):

$$RMSEA = \sqrt{\frac{\chi^2 - df}{[df(N-1)]}} \quad (4)$$

where *N* is the sample size and *df* is the model degrees of freedom. RMSEA is set to zero when χ^2 is less than *df*. Smaller values indicate better fit, with .06 being the proposed cutoff (Hu & Bentler, 1999). RMSEA is typically reported with a 90% confidence interval.

Multilevel MTMM

In social science research hierarchical data structures are routinely encountered, yet go unaccounted for in analysis (e.g. students nested within classes or teachers; families nested within communities; staff members nested within organizations). Scores are clustered under a general unit and at each hierarchy or level the scores tend not be independent (Kline, 2011). It is important to preserve the hierarchical structure of the data.

The independence of observation is the most common assumption of most statistical procedures (Julian, 2001). When the hierarchical (nested) structure of the data is ignored we assume the independence of the outcome is true. In a nested structure (continuing with the example of students nested within classes), within-the-same-class students tend to be more similar than between-classes students by virtue of sharing the same teacher. These observations are therefore not independent because they depend on class. In multilevel analysis this class variance is measured using an intraclass correlation (ICC) index. ICC is expressed as:

$$\rho = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_\varepsilon^2} \quad (5)$$

where σ_c^2 is the within class variance and σ_ε^2 is the error variance. The total variance of the model is $\sigma_c^2 + \sigma_\varepsilon^2$. When the data are truly independent, ICC should equal zero. Additionally, when conducting CFA and its variants, ignoring the hierarchical data structure introduces bias to

chi-square fit statistics, parameter estimates and their standard error estimates (Julian, 2001; Koch, et al., 2015).

Data collected using the ASEBA follows a similar hierarchical structure with children nested within teachers and parents. For example, it is popular with teacher report forms that a single teacher rates multiple students. The rated student scores are nested within a single teacher since teachers are likely to use their own normative understandings about behavior when rating particular students. These normative ratings are likely to vary from teacher to teacher. The purpose of the present study is to examine MTMM where teacher reports on multiple students represent a nested covariate to be accounted for at level two with individual student ratings represented at level one.

Parameters under multilevel MTMM (ML-MTMM) are generated for both group/teacher-level (between group) and individual/student-level (within group) variations. Muthén (1989, 1990) presented a pseudobalanced approach to analyzing multilevel data in SEM. The between group component can be written as $Y_{Bg} = \bar{Y}_g - \bar{Y}$, while the within group component is $Y_{Wgi} = Y_{gi} - \bar{Y}_g$. In these equations *g* represents group and *i* represents individuals. The aggregated cross product matrices are given in (5):

$$\sum_{g=1}^{G}\sum_{i=1}^{n}\left(Y_{gi}-\bar{Y}\right)\left(Y_{gi}-\bar{Y}\right)' =$$

$$n\sum_{g=1}^{G}\left(\bar{Y}_g-\bar{Y}\right)\left(\bar{Y}_g-\bar{Y}\right)' + \sum_{g=1}^{G}\sum_{i=1}^{n}\left(Y_{gi}-\bar{Y}_g\right)\left(Y_{gi}-\bar{Y}_g\right)' \quad (6)$$

where $\bar{Y}$ is the total mean across all teachers and students, $\bar{Y}_g$ is mean of all students of the g^{th} teacher, Y_{gi} is the score of the i^{th} student under the g^{th} teacher, *G* is the total number of teachers, *n*

is the number of students under each teacher (which is assumed to be constant), $N=nG$ is the total students. Muthén (1989, 1990, 1994) showed the sample covariance matrices as S_T, S_W, S_B:

$$S_T = \frac{\sum_{g=1}^{G}\sum_{i=1}^{n}\left(Y_{gi}-\bar{Y}\right)\left(Y_{gi}-\bar{Y}\right)'}{N-G} \tag{7}$$

$$S_B = \frac{n\sum_{g=1}^{G}\left(\bar{Y}_g-\bar{Y}\right)\left(\bar{Y}_g-\bar{Y}\right)'}{G-1} \tag{8}$$

$$S_T = \frac{\sum_{g=1}^{G}\sum_{i=1}^{n}\left(Y_{gi}-\bar{Y}\right)\left(Y_{gi}-\bar{Y}\right)'}{N-1} \tag{9}$$

The between and within groups covariance matrices can also be represented in (9):

$$\sum_T = \sum_B + \sum_W \tag{10}$$

Equation (9) assumes a balanced design yielding results from the maximum likelihood (ML) estimator. Where there are unbalanced designs Muthén (1989, 1990) proposed the use of a Partial Maximum Likelihood or Muthén Maximum Likelihood (MUML) estimator with a scaling parameter (10):

$$c^* = \frac{N^2-\sum_{g=1}^{G}n_g^2}{N(G-1)} \tag{11}$$

Modern SEM software is equipped with the MUML estimator for two-level data, combined with robust estimators for standard errors and test statistics for homogeneity of variance corrections (Hox, 2010). Other methods of analyzing multilevel SEM data have been proposed by Goldstein (1987, 2003) using a multivariate multilevel model and Asparouhov and Muthén (2007) using weighted least squares. Goodness of fit indices, described earlier, is often used to assess model fit in the multilevel SEM framework.

Multilevel CT-C(M-1)

When assessing multilevel MTMM data using CT-C(M-1) models, be it collected from interchangeable or structurally different methods, researchers need to consider five variance components (Carretero-Dios, et al., 2011). *Consistency coefficient* indicates the amount of variance in an item of the non-reference method being explained by the reference method. The consistency coefficient represents convergence of the non-reference and reference methods. A high consistency coefficient therefore, suggests more convergence with the same indicator and thus indicates high convergent validity.

In a structurally different non-reference method, the *method-specificity coefficient* is the portion of variance that is not shared with the reference method. In cases where interchangeable methods are utilized, there are two method-specificity coefficients: the common and unique-specificity coefficients. The *common-specificity coefficient* reflects the degree to which variance of the non-reference method is shared with the other methods but not with the reference method. The *unique-specificity coefficient* shows the proportion of variance of the non-reference method due to the variation of the single non-reference method neither shared with the other non-reference methods nor shared with the reference method. The *reliability coefficient* provides information about the portion of variance in an observed variable being explained by the latent variable, which is not due to measurement error.

Measurement Invariance

Measures are often assessed to determine time, cultural or group similarities (invariances). Measurement invariance is an SEM procedure used to evaluate the magnitude of invariance of a measure. Assessing measurement invariance has been defined by Mellenbergh (1989), Meredith and Millsap (1992), and Meredith (1993); where the central idea behind

measurement invariance is that individual observed scores on a measure are not dependent on measurement occasion or group membership, given their true score value on the measured construct (Chen, 2007; Kline, 2011; Vandenberg & Lance, 2000; Wu, Li & Zumbo, 2007). Measurement invariance is an assumption of many psychometric methods and is a necessary condition to allow for score comparisons across groups (Chen, 2007; Kline, 2011; Vandenberg & Lance, 2000; Wu, et al., 2007).

Measurement invariance constitutes four measurement and two structural steps. Each step in the invariance procedure examines a specified component of the theorized factor structure and assesses the invariance of each component while allowing the other parameters to be freely estimated. The first of the measurement steps is configural invariance. In this step, the researcher assesses the general factor structure of the groups being compared. The second step, factorial invariance, determines whether the groups have the same factor loadings/weights. Scalar invariance is the third step, where the indicator intercepts are assessed between the two groups. In the final step of the measurement invariance process, the residual variances of the groups are tested. This final step is called error variance invariance.

Following the measurement invariance, structural invariance is assessed. The first step determines whether factor variances are the same across groups. The second step examines factor covariances of the groups. By following the steps of both measurement and structural invariance, the researcher is able to relax equivalence constraints in order to determine the parameters that are or are not invariant (Chen, 2007; Vandenberg & Lance, 2000; Wu, et al., 2007).

COMPREHENSIVE REFERENCE LIST

Achenbach, T. M. (1991). *Manual for the child behavior checklist/4-18 and 1991 profile*. Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T. M. (1998). Diagnosis, assessment, taxonomy, ad case formulations. In T. H. Ollendick & M. Hersen (Eds.), *Handbook of child psychopathology* (3rd ed., pp. 63-87). New York: Plenum.

Achenbach, T. M. (2006). As others see us: Clinical and research implications of cross-informant correlations for psychopathology. *Current Directions in Psychological Science*, *15*(2), 94-98.

Achenbach, T. M., Becker, A., Döpfner, M., Heiervang, E., Roessner, V., Steinhausen, H. C., & Rothenberger, A. (2008). Multicultural assessment of child and adolescent psychopathology with ASEBA and SDQ instruments: research findings, applications, and future directions. *Journal of Child Psychology and Psychiatry*, *49*, 251-275.

Achenbach, T. M., Dumenci, L., & Rescorla, L. A. (2002). Ten-year comparisons of problems and competencies for national samples of youth self, parent, and teacher reports. *Journal of emotional and Behavioral disorders*, *10*(4), 194-203.

Achenbach, T. M. & Edelbrock, C. (1983). *Manual for the child behavior checklist/4-18 and revised child behavior profile*. Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T. M., Krukowski, R. A., Dumenci, L., & Ivanova, M. Y. (2005). Assessment of adult psychopathology: Meta-analyses and implications of cross-informant correlations. *Psychological Bulletin*, *131*, 361-382.

Achenbach, T. M., & McConaughy, S. H. (1996). Relations between DSM-IV and empirically based assessment. *School Psychology Review, 25,* 329-342.

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: implications of cross-informant correlations for situational specificity. *Psychological Bulletin*, *101,* 213-232.

Achenbach, T. M., & Rescorla, L. (2001). *Manual for the ASEBA school-age forms & profiles: An integrated system of multi-informant assessment.* Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families.

Achenbach, T. M., & Rescorla, L. A. (2007). *Multicultural supplement to the manual for the ASEBA school-age forms & profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families.

Albrecht, G., Veerman, J. W., Damen, H., & Kroes, G. (2001). The Child Behavior Checklist for group care workers: A study regarding the factor structure. *Journal of Abnormal Child Psychology, 29*(1), 83-89.

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

American Psychiatric Association. (1952). *Diagnostic and statistical manual of mental disorders*. Washington, DC: Author.

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

Asparouhov, T., & Muthen, B. (2007, July). Computationally efficient estimation of multilevel high-dimensional latent variable models. In *Proceedings of the 2007 JSM meeting in Salt Lake City, Utah, Section on Statistics in Epidemiology* (pp. 2531-2535).

Bandura, A. (1986). *Social foundations of thoughts and action: A social cognitive theory.* Englewood Cliffs, NJ: Prentice-Hall.

Barrett, P. (2007). Structural equation modeling: Adjudging model fit. *Personality and Individual Differences, 42,* 815–824.

Bauer, A. M., & Shea, T. M. (1999). *Learners with emotional and behavioral disorders: An introduction*. Upper Saddle River, NJ: Prentice Hall.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238-246.

Bird, H. R. (1996). Epidemiology of childhood disorders in a cross–cultural context. *Journal of Child Psychology and Psychiatry, 37*(1), 35-49.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111,* 1061-1071.

Bower, E. M. (1982). Defining emotional disturbance: Public policy and research. *Psychology in the Schools, 19*(1), 55-60.

Brannick, M. T., & Spector, P. E. (1990). Estimation problems in the block-diagonal model of the multitrait-multimethod matrix. *Applied Psychological Measurement, 14,* 325-339.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: NY, The Guilford Press.

Brown, T. A., & Barlow, D. H. (2009). A proposal for a dimensional classification system based on the shared features of the DSM-IV anxiety and mood disorders: implications for assessment and treatment. *Psychological Assessment, 21,* 256-271.

Browne, M. W. (1984). The decomposition of multitrait-multimethod matrices. *British Journal of Mathematical and Statistical Psychology, 37,* 1-21.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Carretero-Dios, H., Eid, M., & Ruch, W. (2011). Analyzing multitrait-mulitmethod data with multilevel confirmatory factor analysis: An application to the validation of the State-Trait Cheerfulness Inventory. *Journal of Research in Personality, 45,* 153-164.

Castro-Schilo, L., Widaman, K. F., & Grimm, K. J. (2013). Neglect the structure of multitrait-multimethod data at your peril: implications for associations with external variables. *Structural Equation Modeling: A Multidisciplinary Journal, 20*(2), 181-207.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14,* 464-504.

Cole, D. (1987). Utility of confirmatory factor analysis in test validation research. *Journal of Consulting and Clinical Psychology, 55,* 584-594.

Coleman, M. C., & Webber, J. (2002). *Emotional and behavioral disorders: Theory and practice*. Boston, MA: Allyn & Bacon.

Costello, E. J., Egger, H. H., & Angold, A. (2005). One-year research update review: The epidemiology of child and adolescent psychiatric disorders: I. methods and public health burden. *Journal of the American Academy of Child and Adolescent Psychiatry, 44,* 972–986.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Cullinan, D. (2002). Students with emotional and behavioral disorders: An introduction for teachers and other helping professionals. Upper Saddle River, NJ: Merrill Prentice Hall.

Cullinan, D. (2004). Classification and definition of emotional and behavioral disorders. In R. Rutherford Jr., M. Quinn, & S. Mathur (Eds.), *Handbook of research in emotional and behavioral disorders* (pp. 32-53). New York, NY: Guilford Press.

Cullinan, D., & Sabornie, E. J. (2004). Characteristics of emotional disturbance in middle and high school students. *Journal of Emotional and Behavioral Disorders, 12*, 157–167.

Dedrick, R. F., Greenbaum, P. E., Friedman, R. M., Wetherington, C. M., & Knoff, H. M. (1997). Testing the structure of the Child Behavior Checklist/4-18 using confirmatory factor analysis. *Educational and Psychological Measurement, 57*, 306-313.

Dedrick, R. F., Tan, T. X., & Marfo, K. (2008). Factor structure of the Child Behavior Checklist/6-18 in a sample of girls adopted from China. *Psychological Assessment, 20*(1), 70-75.

De Groot, A., Koot, H. M., & Verhulst, F. C. (1994). Cross-cultural generalizability of the Child Behavior Checklist cross-informant syndromes. *Psychological Assessment, 6,* 225-230.

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development, 43*, 121-149.

Eid, M. (2000). A multitrait–multimethod model with minimal assumptions. *Psychometrika, 65,* 241–261.

Eid, M. (2006). Methodological approaches for analyzing multimethod data. In M. Eid, & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 223-230). Washington, DC: American Psychological Association.

Eid, M., Lischetzke, T., & Nussbeck, F. W. (2006). Structural equation model for multitrait-multimethod data. In M. Eid, & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 283-299). Washington, DC: American Psychological Association.

Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait–multimethod models: A multiple-indicator CTC(*M*–1) model. *Psychological Methods, 8*, 38–60.

Eid, M., Nussbeck, F. W., Geiser, C., Cole, D. A., Gollwitzer, M., & Lischetzke, T. (2008). Structural equation modeling of multitrait-multimethod data: Different models for different types of methods. *Psychological Methods*, *13*, 230-253.

van der Ende, J., Verhulst, F. C., & Tiemeier, H. (2012). Agreement of informants on emotional and behavioral problems from childhood to adulthood. *Psychological Assessment*, *24*, 293-300.

Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*, 128-141.

Epanchin, B. (1991). Assessment of social and emotional problems. In J. L. Paul, & B. C. Epanchin (Eds). *Educating emotionally disturbed children and youth: Theories and practices for teachers* (pp. 307-349). New York: Macmillan Publishing Company.

Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research*, 509-529. doi:10.1080/00273170701382864

Federal Register (1981, January 16). Washington, DC: U. S. Government Printing Office.

Flamer, S. (1983). Assessment of the multitrait-multimethod matrix validity of likert scales via confirmatory factor analysis. *Multivariate Behavioral Research, 18,* 275-308.

Forness, S. R., & Knitzer, J. (1992). A new proposed definition and terminology to replace" serious emotional disturbance" in Individuals with Disabilities Education Act. *School Psychology Review*, *21*(1), 12-20.

Furlong, M. J., Morrison, G. M., & Jimerson, S. R. (2004). Externalizing behaviors of aggression and violence and the school context. In R. Rutherford Jr., M. Quinn & S. Mathur (Eds.), *Handbook of research in emotional and behavioral disorders* (pp. 243-261). New York, NY: Guilford Press.

Geiser, C., & Lockhart, G. (2012). A comparison of four approaches to account for method effects in latent state–trait analyses. *Psychological Methods, 17,* 255-283.

Goldstein, H. (1987). *Multilevel models in in educational and social research.* London: Griffin.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: Edward Arnold.

Gomez, R., Vance, A., & Gomez, R. M. (2014). Analysis of the convergent and discriminant validity of the CBCL, TRF, and YSR in a clinic-referred sample. *Journal of Abnormal Child Psychology, 42,* 1413-1425.

Greenbaum, P. E., & Dedrick, R. F. (1998). Hierarchical confirmatory factor analysis of the Child Behavior Checklist/4–18. *Psychological Assessment, 10,* 149-155.

Gresham, F. M., Cook, C. R., Crews, S. D., & Kern, L. (2004). Social skills training for children and youth with emotional and behavioral disorders: Validity considerations and future directions. *Behavioral Disorders*, *30*(1), 32-46.

Gresham, F. M., Elliott, S. N., Cook, C. R., Vance, M. J., & Kettler, R. (2010). Cross-informant agreement for ratings for social skill and problem behavior ratings: An investigation of

the Social Skills Improvement System—Rating Scales. *Psychological Assessment, 22,* 157-166.

Gresham, F. M., & Kern, L. (2004). Internalizing behavior problems in children and adolescents. In R. Rutherford Jr., M. Quinn & S. Mathur (Eds.), *Handbook of research in emotional and behavioral disorders* (pp. 262-281). New York, NY: Guilford Press.

Grigorenko, E., Geiser, C., Slobodskaya, H., & Francis, D. (2010). Cross-informant symptoms from CBCL, TRF, and YSR: Trait and method variance in a normative sample of Russian youths. *Psychological Assessment, 22,* 893-911.

Hox, J. (2010). *Multilevel analysis: Techniques and applications*. New York, NY: Routledge.

Hox, J., & Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling, 8,* 157-174.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55.

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

Individuals with Disabilities Education Improvement Act of 2002, 20 U.S.C. 1401(3)(A)

Ivanova, M. Y., Achenbach, T. M., Dumenci, L., Rescorla, L. A., Almqvist, F., Weintraub, S., ... & Verhulst, F. C. (2007). Testing the 8-syndrome structure of the child behavior checklist in 30 societies. *Journal of Clinical Child and Adolescent Psychology, 36,* 405-417.

Ivanova, M. Y., Achenbach, T. M., Rescorla, L. A., Dumenci, L., Almqvist, F., Bathiche, M., ... & Verhulst, F. C. (2007). Testing the Teacher's Report Form syndromes in 20 societies. *School Psychology Review, 36,* 468-483.

Ivanova, M. Y., Achenbach, T. M., Rescorla, L. A., Dumenci, L., Almqvist, F., Bilenberg, N., ...

& Verhulst, F. C. (2007). The generalizability of the Youth Self-Report syndrome structure in 23 societies. *Journal of Consulting and Clinical Psychology, 75*, 729-738.

Julian, M. W. (2001). The consequences of ignoring multilevel data structures in nonhierarichal covariance modeling. *Structural Equation Modeling, 8,* 325-352.

Kauffman, J. M. (2001). *Characteristics of emotional and behavioral disorders in children and youth* (7th ed.). Upper Saddle River, NJ: Merrill/Prentice Hall.

Kauffman, J. M. (2009, December 23). Emotional/behavioral disorders [Web article]. Retrieved from http://www.education.com/reference/article/emotionalbehavioral-disorders/#A

Kauffman, J. M., & Landrum, T. J. (2009). *Characteristics of emotional and behavioral disorders of children and youth* (9th ed.). Upper Saddle River, NJ: Prentice Hall.

Kenny, D. A. (1976). An empirical application of confirmatory factor analysis to the multitrait–multimethod matrix. *Journal of Experimental Social Psychology, 12,* 247–252.

Kenny, D. A. (2014, October 6). Measuring model fit [Author's Webpage]. Retrieved from http://davidakenny.net/cm/fit.htm.

Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin, 112*, 165-172.

Kenny, D. A., & McCoach, B. (2009). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 333-351. doi:10.1207/S15328007SEM1003_1

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

Koch, T., Schultze, M., Burrus, J., Roberts, R. D., & Eid, M. (2015). A multilevel CFA-MTMM model for nested structurally different methods. *Journal of Educational and Behavioral Statistics, 40*, 477-510.

Ladd, G. W., & Kochenderfer-Ladd, B. (2002). Identifying victims of peer aggression from early to middle childhood: analysis of cross-informant data for concordance, estimation of relational adjustment, prevalence of victimization, and characteristics of identified victims. *Psychological Assessment, 14*(1), 74-76.

Lambert, M. C., Knight, F., Taylor, R., & Achenbach, T. M. (1994). Epidemiology of behavioral and emotional problems among children of Jamaica and the United States: Parent reports for ages 6–11. *Journal of Abnormal Child Psychology, 22*, 113–128.

Lambert, M. C., Lyubansky, M., & Achenbach, T. M. (1998). Behavioral and emotional problems among adolescents of Jamaica and the United States: Parent, teacher, and self-reports for ages 12 to 18. *Journal of Emotional and Behavioral Disorders, 6*, 180–187.

Lance, C. E., Noble, C. L., & Scullen, S. E. (2002). A critique of the correlated trait-correlated method (CTCM) and correlated uniqueness (CU) models for multitrait–multimethod (MTMM) data. *Psychological Methods, 7*, 228–244.

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

Lane, K. L., Barton-Arwood, S. M., Nelson, J. R., & Wehby, J. (2008). Academic performance of students with emotional and behavioral disorders served in a self-contained setting. *Journal of Behavioral Education, 17*(1), 43-62.

Lane, K. L., Pierson, M., & Givner, C. C. (2004). Secondary teachers' views on social competence: Skills essential for success. *Journal of Special Education, 38,* 174–186.

Lane, K. L., Wehby, J. H., & Cooley, C. (2006). Teacher expectations of student's classroom behavior across the grade span: Which social skills are necessary for success? *Exceptional Children, 72,* 153–167.

Lee, S. W., Elliott, J., & Barbour, J. D. (1994). A comparison of cross-informant behavior ratings in school-based diagnosis. *Behavioral Disorders, 19*(2), 87-97.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36,* 437-448.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural equation modeling, 9*(2), 151-173.

Maas, C., Lensvelt-Mulders, G., & Hox, J. (2009). A multilevel multitrait-multimethod analysis. *Methodology, 5*(3), 72-77.

Marsh, H. W. (1989). Confirmatory factor analysis of multitrait–multimethod data: Many problems and a few solutions. *Applied Psychological Measurement, 13,* 335–361.

Marsh, H. W., & Bailey, M. (1991). Confirmatory factor analyses of multitrait-multimethod data: A comparison of alternative models. *Applied Psychological Measurement, 15*(1), 47-70.

Marsh, H. W., & Byrne, B. M. (1993). Confirmatory factor analysis of multitrait–multimethod self-concept data: Between group and within-group invariance constraints. *Multivariate Behavioral Research, 28,* 313–349.

Marsh, H. W., & Grayson, D. (1995). Latent variable models of multitrait–multimethod data. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 177–198). Thousands Oaks, CA: Sage.

Marsh, H. W., & Hocevar, D. (1988). A new, more powerful approach to multitrait–multimethod analyses: Application of second-order confirmatory factor analysis. *Journal of Applied Psychology, 73,* 107–117.

McConaughy, S. H. (1993). Evaluating behavioral and emotional disorders with the CBCL, TRF, and YSR cross-informant scales. *Journal of Emotional & Behavioral Disorders, 1*, 40-52.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13,* 127-143.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58,* 525-543.

Meredith, W. & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement invariance. *Psychometrika, 57,* 289-311.

Merrell, K. W. (2000). Informant reports: Theory and research in using child behavior rating scales in school setting. In E. S. Shapiro & T. R. Kratochwill (Eds.) *Behavioral assessment in schools: Theory, research, and clinical foundations* (2nd ed., pp. 233-256). New York, NY: Guilford Press.

Merrell, K. W. (2008). *Behavioral, social, and emotional assessment of children and adolescents* (3rd ed.). New York, NY: Lawrence Erlbaum Associates.

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

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

Ministry of Education, Youth and Sports (2012). Educational statistical digest of Belize 2011-2012. Belize City: Author.

Mooney, P., Ryan, J. B., Uhing, B. M., Reid, R., & Epstein, M. H. (2005). A review of self-management interventions targeting academic outcomes for students with emotional and behavioral disorders. *Journal of Behavioral Education, 14*, 203-221.

Moreland, A. D., & Dumas, J. E. (2008). Categorical and dimensional approaches to the measurement of disruptive behavior in the preschool years: A meta-analysis. *Clinical Psychology Review, 28*, 1059-1070.

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*, 557-585.

Muthén, B. (1990, June). *Means and covariance structure analysis of hierarchical data*. Paper presented at the meeting of Psychometric Society, Princeton, NJ.

Muthén, B. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research, 22*, 376-398.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Naslund-Hadley, E., Alonzo, H., & Martin, D. (2013). *Challenges and opportunities in the Belize education sector.* Inter-American Development Bank.

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

Nelson, J. R., Benner, G. J., Lane, K., & Smith, B. W. (2004). Academic achievement of K-12 students with emotional and behavioral disorders. *Exceptional Children, 71*(1), 59-73.

Polsgrove, L., & Smith, S. W. (2004). Informed practice in teaching self-control to children with emotional and behavioral disorders. In R. Rutherford Jr., M. Quinn & S. Mathur (Eds.), *Handbook of research in emotional and behavioral disorders* (pp. 399-425). New York, NY: Guilford Press.

Reef, J., van Meurs, I., Verhulst, F. C., & van der Ende, J. (2010). Children's Problems Predict Adults' DSM-IV Disorders Across 24 Years. *Journal of the American Academy of Child & Adolescent Psychiatry*, *49*, 1117-1124.

Reid, R., Gonzalez, J. E., Nordness, P. D., Trout, A., & Epstein, M. H. (2004). A meta-analysis of the academic status of students with emotional/behavioral disturbance. *The Journal of Special Education, 38*(3), 130-143.

Renk, K. (2005). Cross-informant ratings of the behavior of children and adolescents: The "gold standard". *Journal of Child and Family Studies*, *14*, 457-468.

Renk, K., & Phares, V. (2004). Cross-informant ratings of social competence in children and adolescents. *Clinical Psychology Review*, *24*, 239-254.

Reynolds, C., & Kamphaus, R. (2004). *Behavioral assessment system for children* (2nd ed.). Minneapolis, MN: Pearson Assessment.

Richardson, G. A., & Day, N. L. (2000). Epidemiologic considerations. In M. Hersen & R. T. Ammerman (Eds.), *Advanced abnormal child psychology* (2nd Ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Sass, D. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment, 29*, 343-363. doi: 10.1177/0734282911406661.

Schroeder, C. S., & Riddle, D. B. (1991). Behavior theory and practice. In J. L. Paul, & B. C. Epanchin (Eds). *Educating emotionally disturbed children and youth: Theories and practices for teachers* (pp. 148-179). New York: Macmillan Publishing Company.

Shepard, L. A. (1993). Evaluating test validity. *Review of the Research in Education, 19,* 405-450.

Sparrow, S., Cichetti, D., & Balla, D. (2005). *Vineland adaptive behavior scales* (2nd ed.) Minneapolis, MN: Pearson Assessment.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research,* 214-312. doi:10.1207/s15327906mbr2502_4

Synhorst, L. L., Buckley, J. A., Reid, R., Epstein, M. H., & Ryser, G. (2005). Cross informant agreement of the Behavioral and Emotional Rating Scale-(BERS-2) parent and youth rating scales. *Child & Family Behavior Therapy*,*27*(3), 1-11.

Swain, A. J. (1975). *Analysis of parametric structures for variance matrices* (Unpublished doctoral dissertation). University of Adelaide, Australia.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). New York, NY: Allyn and Bacon.

Taylor, T. K., & Biglan, A. (1998). Behavioral family interventions for improving child-rearing: A review for clinicians and policy makers. *Clinical Child and Family Psychology Review, 1*(1), 41–60.

Tinsley, H. E. A., & Weiss, D. J. (2000). Interrater reliability and agreement. In H. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 96-124). San Diego, CA: Academic Press.

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

United Nations, Departement of Economic and Social Affairs, Population Division. (2013). *World population prospects: The 2012 revision, key findings and advance tables.* Working Paper No. ESA/P/WP.227. Retrieved Feb 2, 2014, from http://esa.un.org/wpp/Documentation/pdf/WPP2012_%20KEY%20FINDINGS.pdf

U. S. Department of Education. (2005). 27th annual report to Congress on the implementation of the Individuals with Disabilities Education Act, 2004. Washington, DC: Author.

U. S. Department of Education. (2012). 31st annual report to Congress on the implementation of the Individuals with Disabilities Education Act, 2009. Washington, DC: Author.

Vandenberg, R. J. & Lance, C. E. (2000). A reviews and synthesis of the MI literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-69.

Walker, H. M., Irvin, L. K., Noell, J., & Singer, G. H. S. (1992). A construct score approach to the assessment of social competence: Rationale, technological considerations, and anticipated outcomes. *Behavior Modification, 16*, 448–474.

Webster-Stratton, C., Reid, M. J., & Hammond, M. (2004). Treating children with early-onset conduct problems: Intervention outcomes for parent, child, and teacher training. *Journal of Clinical Child and Adolescent Psychology, 33*, 105-124.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9*(1), 1-26.

Witt, J. C., Vanderheyden, A. M., & Gilbertson, D. (2004). Instruction and classroom management: Prevention and intervention research. In R. Rutherford Jr., M. Quinn & S. Mathur (Eds.), *Handbook of research in emotional and behavioral disorders* (pp. 426-445). New York, NY: Guilford Press.

Wothke, W. (1995). Covariance components analysis of the multitrait–multimethod matrix. In P. E. Shrout (Ed.), *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske* (pp. 125–144). Hillsdale, NJ: Erlbaum.

Wu, A. D., Li, Z. & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice f multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment Research & Evaluation, 12,* Available online: http://pareonline.net/getvn.asp?v=12&n=3.

Yu, C. Y., & Muthen, B. (2002, April). *Evaluation of model fit indices for latent variable models with categorical and continuous outcomes*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.