2020

Psychometric Evaluation of the Eyberg Child Behavior Inventory in an Ethnically Diverse Sample

Elizabeth Machado

PSYCHOMETRIC EVALUATION OF THE EYBERG CHILD BEHAVIOR INVENTORY IN AN ETHNICALLY DIVERSE SAMPLE

by

Elizabeth Machado

A Dissertation presented to the College of Psychology

of Nova Southeastern University

in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy

Nova Southeastern University

2020

Approval Page

This Dissertation was submitted by Elizabeth Machado under the direction of the Chairperson of the Dissertation committed listed below. It was submitted to the College of Psychology and approved in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Clinical Psychology at Nova Southeastern University.

Approved:

June 18th, 2020
Date of Defense

Ralph E. Cash
Ralph E. Cash, Ph.D., ABPP, Chairperson

Ryan Black, Ph.D.

Diana Formoso, Ph.D.

June 25th, 2020
Date of Final Approval

Ralph E. Cash
Ralph E. Cash, Ph.D., ABPP, Chairperson

Statement of Original Work

I have read the Code of Student Conduct and Academic Responsibility as described in the *Student Handbook* of Nova Southeastern University. This dissertation represents my original work, except where I have acknowledged the ideas, words, or material of other authors.

Where another author's ideas have been presented in this dissertation, I have acknowledged the author's ideas by citing them in the required style.

Where another author's words have been presented in this dissertation, I have acknowledged the author's words by using appropriate quotation devices and citations in the required style.

I have obtained permission from the author or publisher—in accordance with the required guidelines—to include any copyrighted material (e.g., tables, figures, survey instruments, large portions of text) in this dissertation manuscript.

Elizabeth Machado
Name

June 1st, 2020
Date

Acknowledgments

I am extremely grateful to those who were instrumental to the completion of my dissertation.

I am most appreciative for the unwavering support provided by my parents throughout this journey. Without their love, encouragement, and generosity my academic pursuits would not have been possible. Mom and Dad, thank you.

To my husband. Thank you for your love, patience, and companionship throughout this endeavor.

To my chairperson, Dr. Cash. Thank you for your willingness, sincerity, and seemingly infinite number of grammar jokes. You were right, we did laugh together.

To Dr. Black. Thank you for the invaluable statistics lessons and challenging me to learn more about Rasch modeling than I ever thought I could.

To Dr. Formoso. Thank you for your kind and constructive approach to mentorship. Your encouragement and expertise were essential to the completion of this study.

Abstract

PSYCHOMETRIC EVALUATION OF THE EYBERG CHILD BEHAVIOR INVENTORY IN AN ETHNICALLY DIVERSE SAMPLE

by

Elizabeth Machado

Nova Southeastern University

2020

This dissertation was designed to confirm the factor structure and to assess the psychometric functioning of the Eyberg Child Behavior Inventory (ECBI) in an ethnically diverse clinical sample using Confirmatory Factor Analysis (CFA) and Rasch modeling. The sample included 221 children and adolescents (72% male and 28% female) whose mothers completed the ECBI. Related to ethnicity, 43.4% of the sample was Hispanic American (HA), 41.2% was European American (EA), 12.2% was African American, and 3.2% identified as "other."

Dimensionality of the ECBI was explored using CFAs and by evaluating model fit criteria. An Andrich Rating Scale Model was employed to assess the rating scale functioning of the ECBI scales. The degree of item invariance across HA and non-HA groups was explored using differential item functioning. Reliability of the scales was assessed using Cronbach's alpha, as well as Rasch-based estimates of reliability.

The results confirmed the superiority of the 3-factor model for the ECBI in an ethnically diverse sample. The 3 scales were found to be unidimensional measures of specific domains of child behavior and their items did not exhibit statistically significant invariance between HA and EA groups. Furthermore, the scales demonstrated acceptable reliability and good convergent and discriminant validity. The findings provided novel empirical support for the cross-cultural use of the ECBI scales and the generalizability of the findings related to the factor structure of the scales to populations with a large HA representation. Lastly, the results revealed that, for the ECBI scales, a 5-category rating scale is optimal for measurement.

Keywords: Hispanic, factor analysis, Rasch modeling, Eyberg Child Behavior Inventory

Table of Contents

List of Tables viii

List of Figures ix

Chapter 1: Statement of the Problem 1

Chapter 2: Review of the Literature 16
- Externalizing Behavior Disorders 16
  - Evidence Based Treatment of Externalizing Behavior Problems 17
  - Evidence Based Assessment of Externalizing Behavior Problems 20
- Cultural Considerations of Externalizing Behavior Disorders 27
  - Hispanic Culture 28
  - Assessment of Externalizing Behavior Problems among Hispanic Youth 32
- Dimensionality in Measurement 39
  - Measuring Dimensionality 41
- The Eyberg Child Behavior Inventory 57
  - The Spanish Version of the Eyberg Child Behavior Inventory 59
  - Standardization of the Eyberg Child Behavior Inventory 61
  - Evidence for a One-Dimensional Measure 63
  - Evidence for a Multi-Dimensional Measure 69
- Present Study 78
  - Hypotheses 81

Chapter 3: Methods 82
- Participants 82
- Sample Characteristics 83
- Measures 83
  - The Eyberg Child Behavior Inventory 83
  - Conners Parent Rating Scale 84
- Analytic Procedure 85
  - Dimensionality 85
  - Rating Scale Functioning 86
  - Model Fit 89
  - Reliability 90
  - Validity 91

Chapter 4: Results 95
- Descriptive Statistics 95
- Hypothesis One 95
- Hypothesis Two 99
  - ODBTA: Rating Scale Functioning 100
  - ODBTA: Dimensionality 105
  - ODBTA: Model Fit 108
  - ODBTA: DIF 112

CPB: Rating Scale Functioning 115
CPB: Dimensionality 119
CPB: Model Fit 122
CPB: DIF 126
IB: Rating Scale Functioning 129
IB: Dimensionality 135
IB: Model Fit 137
IB DIF 141
Validity 143
Hypothesis Three 146

Chapter 5: Discussion 149
Hypotheses 149
A Multidimensional Measure 155
Cross-Cultural Use 157
Limitations of the Study 159
Implications for Future Research 160
Conclusion 162

References 164

Appendices
A. CFA of the Five-point Rating Scale Structure 204
B. Pearson's Product Moment Correlation Coefficients for the Seven-point Rating Scale 207

List of Tables

1 Sensitivity, Specificity, and Predictive Power of the ECBI 11
2 Thresholds for Model Fit Indices in CFA 46
3 Properties of the Eyberg Child Behavior Inventory 59
4 Factor Loadings for Exploratory Factor Analysis 71
5 Descriptive Statistics of the ECBI Total Score 95
6 Model Fit Indices of the One- and Three-Factor Models with Varying Correlated Error Terms for the Seven-point Rating Scale Structure 98
7 Standardized Regression Weights of the Three-Factor Model 99
8 Items of the ODBTA ECBI Scale 101
9 ODBTA Five-point Rating Scale Functioning 103
10 ODBTA Seven-point Rating Scale Functioning 105
11 Results of the PCA of Residuals of the ODBTA Scale 106
12 ODBTA Item Fit Statistics 111
13 ODBTA Person Fit Statistics Summary 112
14 DIF for the ODBTA Scale 114
15 Items of the CPB ECBI Scale 115
16 CPB Seven-point Rating Scale Functioning 117
17 CPB Five-point Rating Scale Functioning 119
18 Results of the PCA of Residuals of the CPB Scale 120
19 CPB Item Fit Statistics 125
20 CPB Person Fit Statistics Summary 126
21 DIF for the CPB Scale 129
22 Items of the IB ECBI Scale 130
23 IB Seven-point Rating Scale Functioning 131
24 IB Five-point Rating Scale Functioning 134
25 Results of the PCA of Residuals of the IB Scale 135
26 IB Item Fit Statistics 140
27 IB Person Fit Statistics Summary 140
28 DIF For the IB Scale 143
29 Guidelines to Describe the Strength of the Correlations 144
30 Pearson's Product Moment Correlation Coefficients for the Five-point Rating Scale 145
31 Separation Coefficients and Reliability Indices 147
A1 Model Fit Indices of the One- and Three-Factor Models with Varying Correlated Error Terms for the Five-point Rating Scale Structure 206
B2 Pearson's Product Moment Correlation Coefficients for the Seven-point Rating Scale 210

List of Figures

1 Category Probability Curve for Item Eleven of the ODBTA Seven-point Scale 103
2 Category Probability Curve for Item Eleven of the ODBTA Five-point Scale 105
3 Contrast Plot of the ODBTA Items' Residual Loadings 108
4 Observed Average Measures Plot of the ODBTA Items 110
5 Category Probability Curve for Item 24 of the CPB Seven-point Scale 117
6 Category Probability Curve for Item 24 of the CPB Five-point Scale 119
7 Contrast Plot of the CPB Items' Residual Loadings 122
8 Observed Average Measures Plot of the CPB Items 124
9 Category Probability Curve for Item 31 of the IB Seven-point Scale 132
10 Category Probability Curve for Item 31 of the IB Five-point Scale 134
11 Contrast Plot of the IB Items' Residual Loadings 136
12 Observed Average Measures Plot of the IB Scale 139

Chapter 1: Statement of the Problem

Approximately 11% to 20% of children in the United States have a behavioral or emotional disorder at some time, with national survey data suggesting increasing prevalence rates (Costello, Mustillo, Ekanli, Keeler, & Angold, 2003; U.S. Department of Health & Human Services, Health Resources & Services Administration, Maternal & Child Health Bureau, 2010). The most common reasons for mental health treatment referral in childhood are externalizing behavior problems (i.e., poor impulse control, aggression, noncompliance) with Attention-Deficit/Hyperactivity Disorder (ADHD) recognized as one of the most prevalent neurodevelopmental disorders of childhood (Centers for Disease Control and Prevention, 2013; Merikangas, Nakamura & Kessler, 2009; Visser et al., 2014). Despite the high prevalence rates, the majority of children in need of mental health services do not receive care (Kataoka, Zhang, & Wells, 2002).

Hispanic and Latino children have higher rates of unmet mental health needs and are less likely to be diagnosed with an externalizing disorder compared to European American children and other minority youth (Alegria, Vallas & Pumariega, 2010; Kataoka, Zhang, & Wells, 2002; Olfson, Moitabai, Sampson, Hwang, & Kessler, 2009). Mental health disparities, such as the lack of standardized assessment and screening procedures across settings, contribute to the under-identification of externalizing behavior problems in minority youth (Mash & Hunsley, 2005; Van de Vijver & Tanzer, 2004; Visser et al., 2014). Because of mental health service disparities among minority youth, nationwide initiatives have sought to alleviate the burden of underserved youth (National Institute on Minority Health and Health Disparities, 2010).

Routine surveillance and screening procedures across settings is recommended to facilitate early identification and treatment of childhood disorders (Beal, 2004; Gall, Pagano, Desmond, Perrin, & Murphy, 2000). Children are often screened for mental health problems with behavior rating scales completed by caregivers in the school or primary care setting (Pagano et al., 2000). In addition, mental health professionals use rating scales for screening, assessment, and treatment purposes (Funderburk, Eyberg, Rich, & Behar, 2003). However, a diagnostic disparity exists among Hispanic youth and the majority of other minority groups (Pumareiga, Rogers, & Rothe, 2005). Many assessment instruments have not proven to be valid for the accurate identification of symptoms and screening of problems across minority youth (Alegria, Vallas, & Pumariega, 2010; Van de Vijver & Tanzer, 2004). Specifically, these authors indicated that measurement equivalence across cultural groups and the potential for response bias are two main concerns when utilizing screening measures and assessment tools.

The use of behavior rating scales, such as the Eyberg Child Behavior Inventory (ECBI; Eyberg & Pincus, 1999), is an efficient and easy method for a variety of professionals to screen for and to assess behavior problems in children (Funderburk, Eyberg, Rich, & Behar, 2003). In particular, the ECBI has been found to be valid and reliable in screening for problematic behaviors and assessing behavior change in children and adolescents (Eyberg & Pincus, 1999). However, the measurement equivalence of the ECBI in culturally diverse samples is relatively unknown, and the dimensionality and factor structure have been scrutinized and criticized due to inconsistent findings (Axberg, Hanse, & Broberg, 2008; Burns & Patterson, 1991; Burns & Patterson, 2000; Colvin, Eyberg, & Adams, 1999; Hukkelberg, 2017; Weis, Lovejoy, & Lundahl, 2005). Similar

to the majority of commonly used rating scales, the ECBI was developed using a largely European American sample with disruptive behavior problems (Eyberg & Robinson, 1983). In addition, recent investigations of the ECBI's psychometric properties and dimensionality have resulted in conflicting evidence relating to the factor structure (Axberg, Hanse, & Broberg, 2008; Burns & Patterson, 1991; Burns & Patterson, 2000) and continue to rely on rather culturally homogenous samples (Colvin, Eyberg, & Adams, 1999; Hukkelberg, 2017; Weis, Lovejoy, & Lundahl, 2005).

Given that the ECBI is widely used in a variety of settings, it is concerning that a consensus relating to its factor structure and dimensionality has not been reached. Additionally, the extant research on the dimensionality and structural invariance of the ECBI is limited by its focus on European American populations. Further investigation of the dimensionality of the ECBI is warranted not only to conclude what is the optimal method of interpreting ECBI scores and to aid in the theoretical understanding of behavior disorders but also to explore the dimensionality and factor structure in a culturally diverse sample.

Generally, "culturally minded" research is necessary because minority groups are overrepresented in the underserved population of children and present with unique mental health care needs, often associated with cultural values and norms (Alegria, Vallas, & Pumariega, 2010; Kataoka, Zhang, & Wells, 2002). For example, Latino immigrant families' perceptions of externalizing disorders may differ from that of American families due to culturally influenced behavior expectations (Monzo & Rueda, 2006). Children of Latino immigrants often participate in most family functions, including adult activities such as grocery shopping, running errands, visiting in the hospital, attending

adult birthday parties, and accompanying family members to medical appointments. In these contexts, Latino children are expected to present with adult-like behavior, and the threshold for what is considered "problematic" behavior differs from cultures in which children are not integrated into as many aspects of adult life.

Evaluation of the ECBI will supplement two areas of research. First, it will add to the literature relating to the validity of the ECBI and the assessment of child externalizing disorders. Second, it will highlight the importance of culturally inclusive research and explore response biases that may be associated with ethnic minorities. By 2050, it is projected that first-generation immigrants will account for 19% of the population in the United States (U.S.), and approximately 18% of the U.S. population will have at least one immigrant parent (Pew Hispanic Center, 2015). As ethnic minority groups continue to become larger percentages of the U.S. population, the mental health disparities among minority groups will become more salient, and the need for culturally competent mental health services will grow (Alegria, Vallas, & Pumariega, 2010).

There are several reasons why minority status has been found to be a relevant factor in the assessment and treatment of childhood externalizing disorders. A main concern is the lack of access to culturally appropriate mental health services which are sensitive to the unique developmental and behavioral expectations of minority groups (Haack & Gerdes, 2011; Pumariega, Rogers, & Rothe, 2005). Screening tools and services that neglect the unique needs of ethnic minorities may not be effective in identifying and treating childhood externalizing problems among minority youth because cultural expectations may influence how individuals experience, express, and address mental health problems (Alegria, Vallas, & Pumariega, 2010; Niec, et al., 2014). In

addition to diagnostic disparities, risk factors associated with externalizing disorders, such as poverty, food insecurity, and contact with juvenile justice systems, disproportionately affect minority youth (Alegria, Vallas, & Pumariega, 2010; Slopen, Fitzmaurice, Williams, & Gilman, 2010).

Mental health disparities among minority youth are apparent across the nation; however, areas with higher rates of immigration, such as the West Coast and the southernmost United States, are especially in need of culturally competent mental health systems due to the presence of larger minority populations (U.S. Census Bureau, 2016). Of note is the Hispanic/Latino population, which has rapidly increased in the United States over the past decade. “Hispanic” or “Latino” refers to an individual of Cuban, Mexican, Puerto Rican, South or Central American, or other Spanish culture of origin (Ennis et al., 2011). The term “Hispanic” will be used throughout this document for consistency. Some areas, such as South Florida, have larger Hispanic populations compared to national averages. According to Census Bureau (2016) data, 28.7 % of the population in Broward County, Florida identifies as Hispanic. In the neighboring county of Miami-Dade, 67.7% of the population is estimated to be Hispanic.

Similar to most underserved groups, Hispanic families experience mental health disparities associated with environmental, societal, and system-related barriers (Alegria, Vallas, & Pumariega, 2010). While Hispanic children are more at risk for the development of externalizing behavior disorders compared to European American families, they are less likely to be identified and to receive interventions (Acevedo-Polakovich, Crider, Kassab, & Gerhart, 2011). Additionally, Hispanic families are more likely to underutilize mental health services (Niec et al., 2014). Underutilization of

mental health care is partly explained by the mismatch between traditional Hispanic values and the mental health services available to Hispanic families. Finally, the dearth of available culturally competent services and empirically supported assessment tools are key contributors to Hispanic mental health disparities (Alegria, Vallas, & Pumariega, 2010).

Disparities in assessment are related to the paucity of empirical evidence supporting the equivalence of assessment tools across racial groups. Specifically, measurement equivalence is a methodological concern often discussed in cross-cultural assessment (Byrne et al., 2009; Van de Vijver & Poortinga, 2005). A lack of measurement equivalence can threaten the comparability of assessment scores as a result of bias. Bias may be related to cultural differences and definitely impacts the construct validity of a measure. Therefore, Byrne and colleagues (2009) discouraged the assumption that meanings of scores are identical across cultural groups. Rather, in order to make meaningful comparisons among scores, there must be evidence that the structural construct is equivalent across groups (Van de Vijver & Tanzer, 2004).

In order to establish equivalence, The Standards for Educational and Psychological Testing (2014) recommends utilizing analytic techniques to identify construct bias as a result of cross-cultural differences. A thorough psychometric evaluation is urged when a measure is intended for use in groups that may be culturally diverse (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). As the population of the United States continues to grow in cultural diversity, cross-cultural measurement equivalence becomes a more salient issue in assessment.

Aside from cultural considerations, the clinical assessment of behavior problems in children is a complicated process due to the complex systems and varied contexts that influence development (Shernoff et al., 2014). Varying models of child development have been used as the foundation of evidence-based assessment (EBA) procedures that take into account the problematic behaviors within the context of the family (Mash & Hunsely, 2005). Due to the complex nature of assessment, the varied settings in which it takes place, and differing professional orientations, a conclusive "gold standard" assessment method for childhood dysfunction has not been identified. However, despite the lack of consensus about assessment strategies, parent-report is agreed to be a core component of the evaluation of childhood problems (Macy, 2012; Shernoff et al., 2014).

Parent or caregiver reports provide primary information regarding child behavior within the framework of the family system (Bruder, 2000; Macy, 2012; Weitzman & Wegner, 2015). Therefore, throughout the assessment process, information relating to symptom severity, frequency, and impairment is often gathered through a combination of parent interview, direct observation, and use of validated behavior-rating scales (Pelham, Fabiano, & Massetti, 2005). In the assessment of behavior disorders in children, providers from multiple disciplines have increasingly come to rely on behavior rating scales as an easy, quick, and reliable method of gathering parent-report information.

The use of behavior rating scales is prevalent across disciplines (Foy, Kelleher, Laraque, & American Academy of Pediatrics Task Force on Mental Health, 2010; Visser, Zablosky, Holbrook, Danielson, & Bitsko, 2015). Currently, the American Academy of Pediatrics Task Force on Mental Health (TFOMH) has a set of practice guidelines delineating screening procedures for symptoms of mental illness and impaired

psychosocial functioning, including behavioral difficulties (Foy, Kelleher, Laraque, & American Academy of Pediatrics Task Force on Mental Health, 2010). The guidelines include routine use of validated screening instruments, such as behavior-rating scales completed by caregivers, for all school-aged children in the primary care setting. In the field of psychology, similar, if not the same, validated rating scales are common in the assessment of externalizing childhood disorders (Funderburk, Eyberg, Rich, & Behar, 2003). For example, a national survey by the Centers for Disease Control and Preventions (CDC) found that behavior-rating scales were used for approximately nine out of ten children assessed for ADHD (Visser, Zablosky, Holbrook, Danielson, & Bitsko, 2015).

Behavior rating scales provide unique advantages compared to other information gathering techniques. First, rating scales can function as broadband measures assessing across a wide range of problems or as focused measures that aid in the assessment of specific behaviors. Second, they are appropriate for use in a variety of settings, including community mental health clinics, medical clinics, hospitals, and schools for screening and assessment purposes (Foy et al., 2010). Third, rating scales that are brief, hand-scored, and psychometrically sound are most desired and utilized across treatment settings (Rich & Eyberg, 2001). Fourth, they allow for the timely collection of information and many can be re-administered in order to monitor treatment progress (Pelham et al., 2005). Last, because young children cannot readily serve as primary informants of their own behaviors, parent-rating scales are especially useful in assessing early childhood functioning. Due to these reasons, behavior rating scales are considered to be the most efficient and widely used methods for screening behavior problems in young children (Funderburk, Eyberg, Rich, & Behar, 2003).

Despite the advantages of parent rating scales, there are notable limitations associated with their use with minority populations. For example, as previously noted, empirical support for use of specific behavior rating scales among Hispanic families is sparse because the majority of mental health research relies on predominately European American samples with limited inclusion of racial/ethnic minorities (Coffey, Javier, & Schrager, 2015; De Los Reyes & Kazdin, 2005; Richters, 1992; Shernoff, Hill, Danis, Leventhal, & Wakschlag, 2014). Most commonly, measurement findings from predominately European American samples are often generalized across populations without ample consideration of cultural differences and response biases. Generalization across groups is concerning due to the cultural differences between non-Hispanic European American and Hispanic children. For example, Hispanic children experience unique stressors related to acculturation, poverty, and language barriers not apparent in most majority populations (Dettlaff & Johnson, 2011). Therefore, assessment tools developed using predominately non-Hispanic European American samples may not be sensitive to the unique mental health needs of Hispanic youth.

One commonly used behavior rating scale that is easily scored, is widely available in a variety of languages, and includes simple to understand items is the ECBI (Eyberg & Pincus, 1999). The ECBI is a 36-item behavior rating scale completed by caregivers to screen for and to assess disruptive behaviors in children and adolescents between two and 16 years of age (Eyberg & Pincus, 1999, Eyberg & Robinson, 1983). Considered a broadband measure of conduct behavior problems in children, the ECBI has empirical support for use as a treatment monitoring tool and is sensitive to assessing behavior change in a variety of cultures (Borrego, Anhalt, Terao, Vargas, & Urquiza, 2006; Burns

& Patterson, 1990; Eyberg, Funderburk, Hembree-Kigin, McNeil, Queriod, & Hood, 2001; Eyberg & Robinson, 1983; Eyberg & Ross, 1978; Nixon, Sweeney, Erickson, & Touyz, 2003; Robinson, Eyberg, & Ross, 1980). The characteristics of the ECBI make it ideal for use in a range of settings for the assessment and treatment of behavior disorders.

The ECBI has been found to be psychometrically sound and valid when used within the recommended populations (Eyberg & Pincus, 1999). It is viewed as a one-dimensional measure with a single factor structure and provides information along two scales (Abrahamse, Junger, Leijten, Lineboom, Boer, & Lindauer, 2015; Colvin, Eyberg, & Adams, 1999; Eyberg & Robinson, 1983; Robinson, Eyberg, & Ross, 1980). Parent responses on the Intensity and Problem scales are summed to provide composite scores with established cut-offs. The scores on the ECBI scales have demonstrated high correlations with the externalizing scale of the Child Behavior Checklist (Achenbach & Rescorla, 2000) and measures of caregiver stress such as the Parenting Stress Index (Abidin, 2012; Boggs, Eyberg, & Reynolds, 1990; Eyberg, Boggs, & Rodriguez, 1992; Haskett, Ahern, Ward, & Allaire, 2006).

Despite the previously noted strengths of the ECBI, it's measurement equivalence across research groups has been questioned in the literature. Evidence relating to the factor structure of the ECBI is inconsistent, and support for a multi-factorial structure has been found (Axberg, Hanse, & Broberg, 2008; Burns & Patterson, 1991; Burns & Patterson, 2000; Hukkelberg, 2017; Weis, Lovejoy, & Lundahl, 2005). Most notably, Burns and Patterson (2000) identified three meaningful factors (i.e., oppositional behavior toward adults, inattentive behavior, and conduct problem behavior) from the intensity scale, which has led to other investigations of the underlying factor structure of

that scale. In fact, Weis et al. (2005) not only found support for the tripartite structure identified by Burns and Patterson (2000) but also found the three factors to have adequate negative predictive power (i.e., ability to rule out particular behavior problems in clinic referred children) and two of the three factors to have adequate positive predictive power (i.e., ability to identify children with significant attention and/or oppositional defiant behavior problems) for externalizing disorders in their sample of young children. Additionally, using two-way contingency analyses, Weis and colleagues assessed the ability of the three component scores to differentiate children with specific externalizing behaviors from children without significant externalizing problems. The results of the analyses by Weis and colleagues are presented in Table 1.

Table 1

Sensitivity, Specificity, and Predictive Power of the ECBI Components

ECBI Indicator	Sensitivity	Specificity	Positive predictive power	Negative predictive power
		Significant attention problems		
Inattentive component	0.77	0.94	0.85	0.90
		Significant oppositional behavior		
Oppositional component	0.75	0.91	0.80	0.82
		Significant conduct problems		
Conduct problem component	0.63	0.94	0.63	0.94

Note. N =115. Sensitivity, specificity, and predictive power statistics reflect each component's ability to differentiate between children with similar behavior problems and clinic-referred children with no significant behavior problems as assessed by a clinician using DSM-IV-TR criteria. Adapted from "Factor Structure and Discriminative Validity of the Eyberg Child Behavior Inventory with Young Children" by R. Weis, M.C Lovejoy, and B.W. Lundahl, 2005, *Journal of Psychopathology and Behavioral Assessment, 27,* 269-278.

Further, Gross et al. (2003) used the tripartite model of the ECBI Intensity Scale in addition to the Intensity Scale total score to evaluate treatment effects of a parent training intervention in a predominately African American and Latino sample. They

determined that there were acceptable alpha reliabilities (α = 0.79, 0.73, and 0.72) for the three individual intensity factors proposed by Burns and Patterson (2000) within their sample. While Gross and colleagues found that parent attitudes related to their child's behavior and discipline strategies improved post-intervention, there were no observed intervention effects on parent-reported child behavior problems for either the total ECBI Intensity Scale score or the three Intensity Scale factors proposed by Burns and Patterson. Notably, the authors alluded to differing cultural values and perceptions as well as the tendency for minority families to underreport child behavior problems as possible explanations for the findings.

Sample and methodological differences in the research make it difficult to draw conclusions relating to the cross-cultural measurement equivalence of the ECBI for two reasons. First, evaluation of the factor structure of the ECBI has involved predominately non-Hispanic European American samples (Axberg et al., 2008; Burns & Patterson, 2000; Colvin et al., 1999; Eyberg & Pincus, 1999; Eyberg & Robinson, 1983; Hukkelberg, 2017; Weis et al., 2005). In fact, just one study including a diverse sample of African American, Hispanic, and non-Hispanic European American participants provided evidence for a single-factor structure (Gross et al., 2007). Second, the factor structure of the ECBI was originally studied using principal components analysis (PCA; Eyberg & Pincus, 1999; Eyberg & Robinson, 1983). Subsequent evaluations of the ECBI have variously used PCA (Burns & Patterson, 1991; Colvin et al., 1999); common factor analysis (Axberg et al., 2008; Burns & Patterson, 2000; Gross et al., 2007; Weis et al., 2005); and, in one study, item-response theory (Abrahamse et al., 2015). Despite the differences in methodology and the inconsistencies in the findings, the ECBI continues to

be widely used in diverse populations as a single-dimensional measure of general disruptive behaviors.

The variability of the factor structure of the ECBI across diverse raters is clinically relevant for several reasons. First, a definitive understanding of the factor structure of the ECBI may increase its utility. For example, results from the three factors of the ECBI could more precisely inform diagnostic formulation and treatment recommendations as part of EBA procedures. In the context of outcome research and intervention evaluation, Burns and Patterson's (2000) tripartite model is argued as more useful because of the ability to parcel out specific domains of behavior change (Axberg et al., 2008; Burns & Patterson, 2000; Weis et al., 2005). Additionally, further investigation of the factor structure may replicate the findings suggesting that the ECBI can be used to differentiate between some externalizing behavior disorders and to identify children likely to have significant attention and/or oppositional defiant behavior difficulties (Weis et al., 2005).

Second, the discrepancies relating to the factor structure of the ECBI call into question the construct validity of the measure. Underrepresentation of ethnic minorities in measure development research can lead to inaccurate assumptions of validity (Haack et al., 2011; Pumariega et al., 2005). The influence of cultural values on item interpretation and response patterns can result in possible response biases or styles, which can alter the factor structure between groups. In order for composite scores to be interpreted and compared appropriately, the latent trait assessed by a scale and the factor structure must be consistent across culturally diverse populations (Van de Vijver & Tanzer, 2004).

Third, the majority of the literature relating to the factor structure of the ECBI relies mainly on a variety of Classical Test Theory (CTT) analytical approaches. CTT approaches are generally "sample specific," suggesting that the findings may, in fact, be true for populations similar to the study sample but may not hold in other populations. Replication of findings using CTT techniques and alternative techniques, such as Item Response Theory (IRT), in culturally diverse samples is warranted to provide further evidence for measurement equivalence. In addition, PCA is an item reduction method at its core, while common factor analysis methods, such as exploratory and confirmatory factor analyses, are used to test theoretical models of latent factors (Conway & Huffcutt, 2003; Schmitt, 2011). Comparing results of PCA and factor analysis methods is common, but somewhat inappropriate, as they are two separate methods. Factor analyses and IRT techniques are appropriate for latent factor evaluation.

In summary, mental health disparities in minority youth further complicate the already complex field of evidence-based assessment of children. Due to the projected trends suggesting significant growth of Hispanic and other minority groups in the U.S., initiatives that address disparities associated with minority status are necessary. In an attempt to close the gap in care, organizations such as the American Psychological Association (APA) have identified mental health disparities as a prominent issue impacting the well-being of minorities (Healthcare Reform: Disparities in Mental Health Status and Mental Healthcare, 2015). As part of health care reform, the APA has called for initiatives focused on the inclusion of culturally diverse groups in research. Culturally inclusive research or cross-cultural research will help providers better understand cultural differences and address disparities associated with cultural diversity.

Since behavior rating scales are widely accepted as routine components in the assessment of children, cross-cultural research of commonly used behavior rating scales is a particularly worthy area for research. Specifically, the ECBI is widely used to obtain parent ratings of problematic behaviors in childhood and is commonly used with families of diverse cultural backgrounds (Borrego, Anhalt, Terao, Vargas & Urquiza, 2006; Burns & Patterson, 1990; Eyberg, Funderburk, Hembree-Kigin, McNeil, Queriod, & Hood, 2001; Eyberg & Robinson, 1983; Eyberg & Ross, 1978; Nixon, Sweeney, Erickson, & Touyz, 2003; Robinson, Eyberg, & Ross, 1980). The reliability and validity of the measure has been demonstrated in several studies. However, the available research on the factor structure of the ECBI is inconsistent and includes predominately non-Hispanic European American samples with limited inclusion of culturally diverse participants. Further, the majority of studies, with some exceptions, utilize CTT techniques, such as PCA, and exploratory and confirmatory factor analyses, which are similar, but not directly comparable.

The paucity of culturally inclusive research samples raises concerns related to the generalizability of the findings to other populations. What is needed is empirical support for the use of the ECBI among culturally diverse populations, such as those with high Hispanic representations, in which the ECBI is already being used. Investigations of the ECBI's factor structure using culturally diverse samples would extend the available literature either to replicate the findings supporting a one-dimensional structure or to provide additional support for the use of the ECBI as a measure of three meaningful dimensions of externalizing behavior disorders.

Chapter 2: Review of the Literature

Externalizing Behavior Disorders

The category of externalizing behavior disorders most often references three distinct types of disruptive behavior. Attention-Deficit/Hyperactivity Disorder (ADHD) is commonly referred to as a disruptive behavior disorder; however, experts conceptualize the disorder as a product of executive functioning deficits (e.g., Barkley, 1997). Oppositional Defiant Disorder (ODD) is a behavior disorder that without intervention is considered a precursor to Conduct Disorder (CD; Burke, Hipwell, & Loeber, 2010; Burke, Loeber, & Birmaher, 2002;). ODD and CD are often paired in the literature, despite evidence supporting a distinction between the two (Bezdjian, Krueger, Derringer, Malone, McGue, & Lacono, 2011).

ADHD is the most commonly diagnosed neurodevelopmental disorder in children (Goldman, Genel, Bezman, & Slanetz, 1998; Merikangas et al., 2009; Visser et al., 2014). The criteria for ADHD as outlined in the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5; American Psychiatric Association, 2015) includes inattention, hyperactivity, and impulsivity symptoms present in at least two settings apparent before age 12. ODD or CD are often co-morbid with ADHD. The ODD criteria include defiant and negativistic behaviors in childhood. CD is characterized by behavior that significantly violates the rights of others and is first apparent in childhood. Although these disorders are no longer listed together in the most recent edition of the DSM-5, all three encompass problematic externalizing behaviors that warrant clinical attention.

Historically, the differentiation between ADHD, ODD, and CD is well-supported (Connor & Doerfler, 2008; Hinshaw, 1987). Still, similarities among the three link them

as externalizing disorders. Therefore, they are best understood as having both shared and unique characteristics reminiscent of a hierarchical model. In an effort to present such a model of externalizing disorders, Bezdjian and colleagues (2011) extracted principal components of ADHD, ODD, and CD criteria from 487 14-year-old males at two time points. Their findings demonstrated that general aspects of externalizing behaviors were at the higher levels of the hierarchy, while more specific features representing individual disorders were at the lower levels. The results supported distinct ADHD, ODD, and CD clustering patterns with subtypes emerging within those clusters (e.g., inattentive and hyperactive/impulsive subtypes of ADHD). The results show that these three externalizing disorders have unique characteristics and share general elements.

The symptoms associated with externalizing behavior disorders largely develop in early childhood and increase the risk of progressing to more severe behavior problems and long-term difficulties lasting into adulthood (Ringel & Strum, 2001; U.S. Public Health Service, 2000; Webster-Stratton & Hammond, 1997). In order to promote early identification, best practice guidelines, including routine screening in pediatric primary care and school settings, have been developed (Chafouleas, Kilgus, & Wallach, 2010; Weitzman & Wegner, 2015). Early identification of disruptive behavior problems followed by appropriate intervention is associated with better long-term outcomes and management of symptoms (Levitt, Saka, Romanelli, & Hoagwood, 2007). Therefore, appropriate screening, assessment, and diagnostic procedures are necessary to ensure that problem behaviors are correctly identified and treated in a timely manner.

Evidence-based treatment of externalizing behavior problems. Given the high incidence of co-morbidity, it is understandable that some of the symptoms associated

with ADHD, ODD, and CD overlap, thus creating a complicated presentation of problematic behavior. Therefore, assessment of such behavior is necessary not only to inform diagnostic formulation, but also to assist in the identification of the most appropriate intervention. A variety of psychosocial interventions have been found efficacious in the treatment of externalizing behavior disorders (Evans, Owens, & Bunford, 2014; Eyberg, Nelson, & Boggs, 2008; Pelham & Fabiano, 2008). While the overarching goal of the majority of externalizing behavior disorder interventions is to reduce disruptive behaviors, each treatment may utilize different techniques to attain that goal. For example, some interventions focus primarily on parenting behaviors, others address the child directly, and alternative programs engage teachers throughout the treatment process. Therefore, only after a thorough understanding of the problem is formulated can an intervention with objectives targeting the relevant characteristics be selected.

Evidence-based treatment (EBT) interventions for ADHD include, but are not limited to, behavior parent-training (BPT), behavior classroom management (BCM), and summer program-based peer interventions (Pelham & Fabiano, 2008). BPT and BCM interventions are often implemented together, and the majority of the research regarding treatment efficacy includes both interventions. Summer Treatment Programs (STPs) are relatively new interventions but have demonstrated positive behavior changes through the use of social skills training, coached group play, and contingency management systems (Pelham, Fabiano, Gnagy, Greiner, & Hoza, 2005; Pelham & Hoza, 1996). Components of BPT, BCM, and STP are often present across interventions and may contribute to the

effects observed in program evaluation studies. However, all are considered efficacious, evidence-based treatments.

The treatment of ODD and CD often focuses on the management of non-compliance, aggression, disruptive classroom behavior, or delinquent behavior (Eyberg et al., 2008). Eyberg and colleagues (2008) identified 16 evidence-based psychosocial treatments for child and adolescent disruptive behavior. Two of the parent-training interventions found to be efficacious in the treatment of ODD were the Incredible Years Parent Training (Reid, Webster-Stratton, & Hammond, 2003) and Parent-Child Interaction Therapy (Brinkmeyer & Eyberg, 2003). Skills training programs, such as the Problem-Solving Skills Training (Kazdin, 2003), have also been found to be evidence-based interventions for disruptive behavior disorders. Evidence-based interventions for children and adolescents with more serious antisocial and delinquent behaviors include Multisystemic Therapy (Henggeler & Lee, 2003) and Multidimensional Treatment Foster Care (Chamberlain & Smith, 2003).

Assessment is not only an ongoing component of EBT, it is often the first step in the treatment process. In order to select the most appropriate EBT, assessment of the presenting concern is necessary. For example, a child who is experiencing functional impairment primarily due to symptoms of inattention would likely benefit from an intervention that is different from one that would be warranted for a child whose majority of concerns are associated with non-compliance and defiance. In addition to the selection of an EBT, assessment facilitates early identification, intervention monitoring, and treatment efficacy (Mash & Hunsley, 2005).

Evidence-based assessment of externalizing behavior problems. The primary goals of assessment of child dysfunction include discriminating abnormal functioning from normal functioning, understanding impairment, and identifying strengths and weaknesses for the purposes of diagnostic clarification, case conceptualization, treatment planning, and/or the evaluation of progress (Achenbach, 2017; Mash & Hunsley, 2005). Historically, behavior problems in childhood have presented a unique assessment challenge. This is due, in part, to the diverse settings in which the screening and assessing of problems occurs, as well as the dearth of evidence-based assessment (EBA) guidelines to compliment EBTs (Achenbach & Ruffle, 2000; Mash & Hunsley, 2005). For example, given the evolving nature of child psychology, children may be referred to diverse settings/professionals such as community mental health clinics, pediatric primary care clinics, or school psychologists for evaluation. There has also been a shift from lengthy and generic test batteries to the use of disorder-specific and brief batteries that can be more easily integrated into treatment services (Mash & Hunsley, 2005; 2007). As the theory of EBA develops, clear and consistent guidelines to standardize EBA practices across settings are necessary to bridging the gap between EBTs and EBAs.

In order to facilitate the shift toward EBA practices, experts in the field have explored what EBA methods entail. For example, in a special section of the *Journal of Clinical Child and Adolescent Psychology,* Mash and Hunsley (2005) discussed the complexities of and challenges inherent in EBA. Additionally, they identified various dimensions to consider when deciding whether a measure is evidence-based, as well as what factors contribute to EBA. The importance of utilizing assessment tools that have published standardization, reliability, and validity information; have population-based

norms; can be easily used by community providers to inform treatment planning; and can be re-administered for treatment monitoring were highlighted as essential components of EBA practices. In addition, in a review of child assessment literature, Kazdin (2005) outlined five additional themes of EBA. These include, but are not limited to, the advantage of utilizing multiple informants and sources of information to obtain varied perspectives on a problem and the consideration of influences on performance, such as ethnicity, when interpreting scores.

A main barrier associated with adhering to the EBA recommendations proposed by Mash and Hunsley (2005) and Kazdin (2005) is the paucity of clear criteria and standards to evaluate an assessment method and to deem it "evidence-based." This barrier is applicable to most of the commonly used assessment methods, including caregiver interviews, structured parent report forms, and self-rating methods (Mash & Hunsley, 2007). However, when selecting an assessment procedure, it is generally assumed that a combination of varied assessment modalities is best (Achenbach, 2017).

There are two primary reasons multi-method assessments with parent or caregiver involvement are widely recognized as crucial to the evaluation of behavior disorders in children (Achenbach, 2017; Macy, 2012; Shernoff et al., 2014). First, multi-method assessment procedures allow clinicians to gain insight into behavior in a variety of settings, to assess strengths and weaknesses, and to understand the perspectives and relationships of different informants (Achenbach, 2017; Salbach-Andrae, Lenz, & Lehmkuhl, 2009). Second, given that behavior problems are often complex and that co-morbid disorders are frequently present, gathering information from different domains improves the clinician's understanding of the problematic behaviors within the varied

contexts in which they occur. For these purposes, behavior rating scales are the most widely used structured method for garnering standardized information from parents or caregivers (Gresham, Elliot, Cook, Vance, & Kettler, 2010; Mash & Hunsley, 2007).

Rating scales. In the assessment of children, depending on their age, informants often include the child's parents or caregivers and teachers (Achenbach, 2017). While clinical interview is the most commonly used assessment procedure, interview information is often integrated with other types of data gathered from parent rating scales. When completing rating scales, informants are asked to make scaled judgments relating to the presence or the degree of impairment associated with a particular behavior (Mash & Hunsley, 2007). Rating scales offer a standardized, cost effective, and timely method for gathering information (Mash & Hunsley, 2007; Salbach-Andrae et al., 2009). Further, the scales can be used by a variety of professionals in diverse settings. For example, in an analysis of data collected through the 2014 National Survey of the Diagnosis and Treatment of ADHD, Visser and colleagues (2014) found that the majority of diagnoses of ADHD made by a primary care physician, psychiatrist, or psychologist included the use of one or more behavior rating scale(s) or checklist(s) in addition to a parent interview.

Compared to other methods of child assessment, such as direct observation, rating scales demonstrate unique strengths. Considered to be an *indirect* measure of behavior, rating scales provide insight into the retrospective occurrence of behaviors, while direct observation is used to measure behaviors as they occur (Gresham & Lambros, 1998). Although both methods of measurement serve important roles in assessment, behavior rating scales have several advantages (Gresham & Lambros, 1998; Mash & Hunsley,

2007). First, such scales allow for the collection of quantifiable data supported by pre-established reliability and validity. Second, based on the purpose of the assessment, they can be used to assess a broad range of behavior or a narrowly targeted behavior in a timely manner. Third, multiple informants can be used to assess behavior from various perspectives and settings, as well as at different points in treatment. Fourth, the use of validated rating scales allows for the comparison of results to normative data in order to understand better the severity of the behaviors (Gresham & Elliot, 2008; McConaughy & Ritter, 2005).

Rating scales have utility at every stage of evidence-based practice, including screening, assessment, treatment, and outcome (Achenbach, 2017). For example, routine screening for childhood mental health difficulties can facilitate early identification and timely referral for services. This is especially pertinent, as early intervention has been found to reduce the risk of ongoing disruptive behaviors in adolescence and adulthood (Guralnick, 2011; Levitt et al., 2007). Further, rating scales assist in identifying a specific problem behavior or area of deficit, and this identification can contribute to selecting an appropriate EBT (Mash & Hunsley, 2007). In addition to their utility as screening instruments, rating scales which have sufficient test-retest reliability are often used to monitor treatment progress.

Beyond clinical utility, rating scales have proven especially useful in meeting the needs of the changing nature of mental health care and assessment. As was previously noted, assessment and treatment of childhood dysfunction is no longer limited to traditional mental health settings. Because the field of clinical psychology and service reimbursement models continue to evolve, the need for cost-effective and well-validated

measures of treatment efficacy and of behavior grows (Brestan, Jacobs, Rayfield, & Eyberg, 1999; Plante, Couchman, & Diaz, 1995). Therefore, assessment tools that are brief, gather multiple informant information in a timely manner, and are cost-efficient can be invaluable to the assessment of child behavior across treatment settings (Kamphaus, Petoskey, & Rowe, 2000; Mash & Hunsely, 2007).

A main concern relating to the use of parent behavior rating scales is the extent of agreement among informants (De Los Reyes, 2011; Mash & Hunsely, 2007). Accepting modest agreement among multiple raters of child functioning has been the long-standing norm, with little guidance provided as to how to improve agreement (Achenbach et al., 1987; Sawyer, Baghurst, & Clark, 1992). In a review of the available literature, Achenbach and colleagues (1987) found the inter-rater agreement among parents, teachers, and mental health workers to be statistically weak (i.e., $r = 0.20$). More recent reviews of the literature continue to reference cross-informant report discrepancies as a challenge when assessing child psychopathology, suggesting that little has changed since Achenbach and colleagues' meta-analysis (De Los Reyes, 2011; De Los Reyes & Kazdin, 2005; Rescorla, et al., 2013; Salbach-Andrae et al., 2009). Nevertheless, the use of multiple informants is still considered a "best practice" standard in evidence-based assessment. (Dirks, De Los Reyes, Briggs-Gowan, Cella, & Wakschlag, 2012; Hunsley & Mash, 2007).

Although the aim of gathering collateral information is to establish convergence among raters or settings, discrepancies among informant responses can yield valuable information (Achenbach, McConaughy, & Howell, 1987; Mascendaro, Herman, & Webster-Stratton, 2012). For example, if a measure is psychometrically sound, weak

levels of cross-informant agreement may not be a challenge but rather a tool that can be used in conceptualization and treatment (Achenbach, 2017; Poston & Hanson, 2010). Specifically, discrepant profiles of caregiver reports can be used to provide feedback to caregivers about their perceptions of their child's behavior and to increase their understanding of child behavior and management.

Multiple informant report convergence continues to be a well-researched area of interest due to limited understanding of the conditions under which these perceptions agree or diverge. Several factors, including informant psychological symptoms (e.g., maternal mental health concerns), relationship dynamics among informants (e.g., marital discord, divorce), and parental acceptance of the child (e.g., parenting satisfaction), have been examined in order to understand report discrepancies better (Kolko & Kazdin, 1993; Treutler & Epkins, 2003) In addition, the problem type and the informant's race/ethnicity have been identified as important factors contributing to degree of informant agreement (Achenbach et al., 1987; Duhig, Renk, Epstein, & Phares, 2000; Mascendaro et al., 2012; Treutler & Epkins, 2003; Youngstrom et al., 2000).

Relating to problem type, higher levels of convergence have been found for externalizing compared to internalizing behavior ratings (Achenbach et al., 1987; Duhig, Renk, Epstein, & Phares, 2000) with some exceptions (Kolko & Kazdin, 1993). Race/ethnicity has been found to be a second significant factor associated with informant agreement. For example, Youngstrom, Loeber, and Stouthamer-Loeber (2000) found the overall discrepancies among informants in their sample to be consistent with the findings of Achenbach and colleagues (1987) as well as De Los Reyes and Kazdin (2005). However, Youngstrom et al. (2000) also found race to be associated with higher levels of

divergence between teacher and parent reports of externalizing problem, as well as between teacher and youth ratings of externalizing problems. Specifically, teachers reported higher levels of externalizing problems (average of 3.2 points higher) for African American males than for European American males, compared to caregiver and self-report. These findings are consistent with other results suggesting that teachers perceive African American children as having more disruptive behaviors than European American children (Pigott & Cowen, 2000). Additionally, parenting stress and caregiver depressive symptoms have been found to be a predicting factor of the variance associated with report discrepancies between informants (Van der Oord, Prins, Oosterlaan, & Emmelkamp, 2005; Youngstrom, Loeber, & Stouthamer-Loeber, 2000).

A noticeable gap in the research on convergence is the dearth of multiple caregiver reports, despite the EBA recommendation that information be gathered from all primary caregivers. Most often, parental reports are obtained from the child's mother and paternal reports are absent. For example, 91% of the data analyzed by Youngstrom et al. (2000) relied on maternal reports. Nevertheless, in studies that have included multiple caregiver's reports, similar weak levels of cross-informant agreement were observed regarding both externalizing and internalizing disorders (Achenbach et al., 1987; Duhig et al., 2000; Mascendaro et al., 2012; Treutler & Epkins, 2003).

In general, clinicians seek out maternal reports over those of father's because they are often considered to be the more accurate accounts of behavior in children (Phares, 1992; Phares, 1997; Phares, Lopez, Fields, Kamboukas, & Duhig, 2005). The reasoning for this assumption most likely reflects past societal attitudes that often characterized mothers as primary caregivers who are, therefore, more familiar with child behavior.

However, relying on the report of only one parent/caregiver can have implications for data interpretation and determinations of clinical significance because of varying factors, such as parenting stress and informant mental health, that can influence such reports (Bingham, Loukas, Fitzgerald, & Zucker, 2003; Hulbert, Gdowski, and Lachar, 1986). Therefore, in order to conceptualize problematic behaviors most effectively, understanding the factors that may influence informant responses, such as parenting stress and parent psychopathology, is recommended.

In sum, Achenbach (2017) posited that EBT of externalizing disorders cannot be optimally done without EBA. Specifically, he concluded that EBA informs "whether to treat, who to treat, what to treat, how to treat, and how much to treat" (Achenbach, 2017, p. 161). To this end, assessment procedures, combining a variety of assessment media, including rating scales completed by multiple informants, are recommended in the assessment of externalizing disorders. However, in order for a rating scale to be appropriately utilized, there must exist evidence for its use in the target population. Critically reviewing the test development procedure and normative populations shows whether the measure will yield reliable information to the clinician. Additionally, understanding the underlying construct as well as any latent factors of a measure further informs whether that measure will be appropriately used.

Cultural Considerations of Externalizing Behavior Disorders

Not all children who act out are equally likely to be diagnosed with and to receive treatment for externalizing behavior disorders and diagnostic disparities, such as the underdiagnosis of externalizing behavior disorders, exist. For example, females, as well as ethnic minority children, are diagnosed with ADHD at lower rates compared to their

European American and male counterparts (Morgan, Staff, Hillemeier, Farkas & Maczuga, 2013; Schnieder & Eisenberg, 2006). In particular, while Hispanic children are less likely to be diagnosed with ADHD compared to non-Hispanic European American children, they are not less likely to display ADHD-related behaviors. Further, Hispanic children diagnosed with a behavioral health disorder, including ADHD, are also less likely to receive quality intervention and are more at risk for premature treatment drop-out (Morgan, Staff, Hillemeier, Farkas & Maczuga, 2013; Olfson, Moitabai, Sampson, Hwang, & Kessler, 2009; U.S. Surgeon General Report, 2001). Due to the unmet mental health needs and growing Hispanic population, researchers have begun to test theoretical models in an effort to better understand the mental health disparities in Hispanic populations.

Hispanic culture. "Culture" is a general term that encompasses values, norms, and experiences of a group of people (Canino & Guarnaccia, 1997). It is a milieu that can be shared by a large group of individuals or developed within a small group as a result of unique life experiences. While there is a tendency to equate culture with ethnic groups, this may oversimplify the concept of cultural units (Harwood, Schoelmerich, Ventura-Cook, Schulze, & Wilson, 1996). Although shared ethnicity can represent a group's commonalities, variations within that group demand appreciation. Therefore, a more fluid understanding of culture involves the recognition that individuals often belong to several cultural groups at varying levels of inclusion.

The term "Hispanic" refers to a person's ethnicity and heritage rather than to race (U.S. Census Bureau, 2016). Hispanic culture is diverse and not well-defined by a universal label if the variations within the group are overlooked (Canino & Guarnaccia,

1997). Dimensions of Hispanic culture can vary by nation of origin, migration, and relationship to the United States. Although consideration of the variations within Hispanic culture is ideal, Hispanic heritage has been found to encompass common characteristics related to socialization, familial relations, and child-rearing practices that apply to the majority of Hispanic individuals (Canino & Guarnaccia, 1997; Harwood et al., 1996).

Hispanic culture and youth mental health. Cultural distinctions of socialization, such as being *sociocentric* or *egocentric*, are widely accepted as methods by which to understand developmental differences across cultures (Harwood, Handwerker, Schoelmerich, & Leyendecker, 2001; Hollan, 1992). Sociocentricism is a cultural dimension relating to the development of an individual's identity within the context of a larger group. In sociocentric cultures, identity is developed from a group or the extended family, status within the group, and the group's status in the larger society. In *egocentric* cultures, a person's identity is relatively independent of the group and being dependent on others is looked down upon. Rather than a binary concept, socialization is better understood as being on a spectrum and is associated with child-rearing practices and parental expectations of conduct.

Generally, Hispanic culture is more sociocentric compared to North American cultures, and this sociocentrism is largely in opposition to the North American emphasis on autonomy (Harwood et al., 1996). Developmentally, Hispanic children are often socialized to value connections to others and to integrate with social networks (Canino & Guarnaccia, 1997). Indulging children is used as a way to build the parent-child relationship and is a common parenting practice within Hispanic culture. In general,

Hispanic Americans, including those from Mexico, Central America, and Cuba, demonstrate a strong attachment with family members and have powerful feelings of loyalty to their families (Sabogal, Marin, & Otero-Sabogal, 1987). These culturally specific values are referred to as *familismo.*

Relating to expectations of conduct, Hispanic families value different characteristics of behavior compared to non-Hispanic European American families. For example, Hispanic children are encouraged to be calm, well-mannered, and respectful toward adults above all else (Harwood et al.1996; Harwood et al., 2001). The emphasis on obedience and consideration in Hispanic culture is referred to as *respeto* (Calzada, Fernandez, & Cortes, 2010). In addition, children are expected to integrate easily into the extended family network and to maintain close familial relationships. These behavioral and temperamental expectations may make Hispanic families more sensitive to deviations, such as non-compliance, hyperactivity, and aggression, even when these behaviors may be considered developmentally appropriate in other families. Further, managing misbehavior differs considerably among Hispanic families. Physical means of behavior management are widely acceptable, albeit as a last resort after verbal attempts are unsuccessful (Calzada, Fernandez, & Cortes, 2010; Monzó & Rueda, 2006).

Acculturation is an important concept in the majority of Hispanic cultures in the United States in addition to unique socialization and child rearing practices (Bernal & Sáez -Santiago, 2006; Dinh, Roose, Tein, & Lopez, 2002). Acculturation is the process by which a person adapts to a new living environment and integrates the norms and values of the new setting (Abraído-Lanza, Armbrister, Flórez, & Aguirre, 2006). As the individual acculturates, acculturative stress can result from changes in the family system,

conflicting cultural values, and language barriers (Bernal & Sáez -Santiago, 2006). Moreover, Hispanic children and adolescents may experience conflict as a result of inconsistency between behavioral expectations at home and those they observe in the broader environment (Canino & Guarnaccia, 1997; Dinh et al., 2002). The influences of acculturation can be observed in parent-child relationships, family dynamics, and social relationships. Specifically, a higher level of acculturative stress is a risk factor for externalizing disorders and depressive symptomatology in Hispanic children and adolescents (Cano et al., 2015; Dinh et al., 2002).

In addition to individual and family factors, environmental factors such as current geographical region are possible links to acculturative stress and mental health outcomes among Hispanic youth (Cano et al., 2015; Lawton & Gerdes, 2014; Yabiku, Kulis, Marsiglia, Lewin, Nieri, & Hussaini, 2007). For example, Yabiku and colleagues (2007), found that for Hispanic youth, residing in an area with a highly concentrated Hispanic population was protective against substance use, while living in a predominately non-immigrant area was a risk factor for alcohol and marijuana use. Further, immigrants living in environments with high immigrant populations experience lower acculturative stress and higher accessibility to culturally competent services (Lawton & Gerdes, 2014). In addition, experiences such as discrimination and socioeconomic factors vary by region. Therefore, assuming that the experiences of all Hispanic families across the United States are similar is an oversimplification of a complex issue. Specifically, Cano and colleagues (2015) found that higher reports of acculturative stress predicted increased depressive symptoms among Hispanic participants in Miami but not Los Angeles. These

results show that, when developing culturally tailored services for Hispanic populations, differences associated with geographical location are important considerations.

The growing body of research related to mental health issues among Hispanic populations and other ethnic minorities suggests that unique cultural values and acculturative stressors may limit the effectiveness of mental health services that have been successful in predominately non-Hispanic European American populations (Cano et al., 2015; Dinh et al., 2002; Monzó & Rueda, 2006). In order to address the Hispanic mental health disparity, culturally and geographically tailored services are necessary, due to the variations in experiences within Hispanic American groups (Geisinger, 1994; Lawton & Gerdes, 2014). Screening tools and assessment methods that are valid and reliable across groups are especially important in order to improve identification rates within underserved populations.

Assessment of externalizing behavior problems among Hispanic youth. General factors, such as misdiagnosis, barriers to access to mental health services, and lack of culturally sensitive validated assessment measures, contribute to the mental health disparities among ethnic minority groups (Pumariega et al., 2005). The fact that there are many cultural factors and values unique to Hispanic populations suggests that culturally specific initiatives are needed to improve the effectiveness of mental health care for this group (Bridges, Andrews, Villalobos, Pastrana, Cavell, & Gomez, 2014; Escobar, Burnam, Karno, Forsythe, & Golding, 1987; McCabe & Yeh, 2009; McCabe, Yeh, Garland, Lau, & Chavez, 2005). For example, Hispanic individuals are more likely to access traditional medical services instead of mental health services. Further, Hispanic individuals are more likely to express somatic complaints in response to psychological

distress. Additionally, child rearing practices and culturally specific expectations of child behavior influence how parents perceive, manage, and address child behavior problems (Halguenseth et al. 2006). Therefore, screening tools and assessment measures used to identify behavior problems in North American families may not be useful in detecting problematic behaviors in Hispanic families.

Common methods of assessing externalizing behavior problems, such as parent interviews and questionnaires, may not accurately identify symptoms in children from Hispanic families. In a critique of the literature related to functional impairment and ADHD, Haack and Gerdes (2011) identified several factors that may explain why symptom report, a common practice in the assessment of ADHD, may not be a reliable method when considering an ADHD diagnosis in Hispanic individuals. Although Haack and Gerdes (2011) applied these factors to the assessment of functional impairment in ADHD, they can be more generally used to understand better the unique challenges to mental health assessment among Hispanic families.

First, the collective values typically observed in Hispanic families, i.e., personalismo and familismo, may facilitate a more accepting and understanding view of child behavior (Borrego et al., 2006; Canino & Guarnaccia, 1997; Halgunseth, Ispa, & Rudy, 2006). Hispanic parents who maintain collectivistic cultural values may be less likely to rate externalizing symptoms as problematic (Schmitz & Velez, 2003). Second, validated assessment measures may not be available in Spanish and may not take into account the attitudes, beliefs, values, and expectations that may differ from what is observed in non-Hispanic European American families (Padilla & Medina, 2001; Rothe, 2005). Some measures have been translated to Spanish to help address this barrier;

however, translation does not ensure that a measure demonstrates the same psychometric properties in populations of different cultures.

"Cultural adaptation" is an effort to modify measurement tools and interventions for use in different cultures. Cultural adaptation goes beyond simple translation of instruments and incorporates culture-specific modifications which manage issues of culture that may interfere with response patterns or treatment efficacy (Matos, Torres, Santiago, Jurado, & Rodriguez, 2006; Niec, et al., 2014). Cultural adaptations to evidenced-based treatments have demonstrated positive outcomes. For example, culturally modified versions of Cognitive Behavior Therapy, Parent Management Training, and Parent-Child Interaction Therapy have been found effective in treating Puerto Rican and Mexican adolescents (Martinez & Eddy, 2005; Matos, Torres, Santiago, Jurado, & Rodriguez, 2006; McCabe & Yeh, 2009; Rosselló & Bernal, 1999). However, less attention has been given to the cultural adaptation of evidence-based assessment measures compared to interventions. The majority of the focus is often on translating the language of the measure, which neglects the cultural component.

Cultural considerations for evidence-based assessment. Generally, there is consistent effort made by clinicians to utilize validated assessment tools that undergo a series of psychometric analyses to ensure evidence-based practice. However, Kazdin (2005) posits that psychometric evaluation is a never-ending process because no number of studies can exhaust one kind of validity or provide normative data from all possible samples. Therefore, assessment practices among culturally diverse populations should be approached with caution, because assuming that psychometric findings that are true for one culture hold true for all cultures can lead to misunderstandings and inaccurate

assessment outcomes (Van de Vijver & Poortinga, 2005). In addition, given that the majority of validation studies have a limited inclusion of ethnic minorities, the possibility that a measure may not function comparably among different ethnic or demographic groups is likely (Mash & Hunsley, 2005). Therefore, supporters of evidence-based assessment recommend that clinicians look beyond the stated validity and reliability and keep in mind the gender, ethnicity, and age of the rater (Achenbach, 2017; Kazdin, 2005).

In an effort to address the obstacles to valid cross-cultural assessment, Van de Vijver and Poortinga (2005) proposed a classification system to standardize the evaluation of measures across cultures. The authors identified two levels of equivalency, i.e., structural and measurement, needed before a measure can be used cross-culturally. Structural equivalence refers to the extent to which the meaning and dimension of a construct is similar across cultural groups (Byrne et al., 2009; Van de Vijver & Poortinga, 2005). Measurement equivalence is the extent to which the item content and psychometric properties are comparable across groups. In order to make meaningful comparisons of scores among culturally diverse groups, there should exist factorial invariance and item equivalence across groups. This is especially true for measures that have been translated or adapted for use among diverse populations. An analytic method suggested by Van de Vijver and Poortinga (2005) to examine the factorial invariance across cultural populations involves exploratory or confirmatory factor analyses. In addition to factor analysis, modern test theory analyses have become increasingly popular in cross-cultural research (Byrne et al., 2009).

The assessment standards proposed by van de Vijver and Poortinga (2005) and the elements of EBA are comparable. For example, Achenbach (2017) recognized that

the majority of mental health assessment research focused on a handful of rather similar cultures. Therefore, he moved to expand the scope of EBA methods advocated by Kazdin (2005) and Mash and Hunsley (2005) by emphasizing the need for assessment practices that are both evidence-based and appropriate for diverse populations. Specifically, Achenbach stressed that testing the applicability of measures among cultural groups before clinical use is necessary in EBA.

In order to demonstrate how a measure can be evaluated for cross-cultural use, Achenbach (2017) presented analytic findings from CBCL data collected from more than fifty cultures. Results of confirmatory factor analyses (CFA) of the CBCL syndrome scales were similar to the syndromes identified in the original Anglophone samples from the United States (Achenbach, 2017; Achenbach & Rescorla, 2000; Achenbach & Rescorla, 2001). However, differences among mean scale scores between cultures were found. The findings suggest the CBCL performs similarly among diverse cultures; however, the development of various sets of norms for clinical use was warranted (Achenbach, 2017; Achenbach & Rescorla, 2015).

Achenbach's (2017) review not only highlights the need for EBA practices that are appropriate across cultures, but it also provides a methodological framework as to how researchers can evaluate the applicability of a measure to diverse populations. The methodology Achenbach described is comparable to the recommendations made by Van de Vijver and Poortinga (2005). Although alternative terminology is used, both proposals emphasize the importance of the factorial invariance of a measure across cultures and recommend factor analytic techniques to test the assumption.

Cultural considerations of rating scales. Rating scales are especially vulnerable to sources of error and non-equivalence, given the wide range of settings, regions, and populations in which they are used (Achenbach, et al., 2008; Byrne et al., 2009). Systematic errors, including halo effects, resulting from respondent tendencies to lean toward certain sets or items call for caution when interpreting results. For example, cultural values can contribute to construct biases and differences in dimensional structures. Further, rating scales can fail at capturing the respondent's interpretation of items, which may lead to response biases. Therefore, in order to reduce the potential for biased results and to avoid inaccurate conclusions about a child's mental health, a measure must be valid for use within the specified population.

An example of how differing cultural values can influence the equivalence of a measure can be found in a review of the functioning of the Achenbach System of Empirically Based Assessment (ASEBA; Achenbach, 2009) and the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997). Achenbach and colleagues (2008) investigated the reliability and validity of those two measures among multi-cultural populations. In their review, the authors included an abundance of psychometric findings for each measure from more than 30 societies to demonstrate the possible variances of scores between cultures. Sufficient evidence was found to conclude that the ASEBA and SDQ were appropriate for use among diverse populations. However, the authors found some variability in model fit and evidence that alternative factor models would fit the data. In order to explain the factorial invariance across groups, cultural and societal views were identified as possible sources of variance. Specifically, how informants viewed

behavior and responded to items likely influenced the variability in model fit and contributed to the nonequivalence.

Method bias arising from unique response styles is a particularly relevant concern when using rating scales among Hispanic populations. In addition to the varying cultural values, Hispanic individuals tend to demonstrate an extreme response pattern on Likert-type scales (Bachman, O'Malley, & Freedman-Doan, 2010; Batchelor & Miao, 2016; Hui & Triandis, 1989). The tendency to select items at either extreme of a scale is thought to be associated with the value placed on sincerity within Hispanic cultures. Extreme response style (ERS) was first identified by Cronbach (1946) as the observed pattern of some individuals consistently to use the extreme ends on response scales. Similar to the conclusions reached by Achenbach et al. (2008), response patterns, such as ERS, can influence the factor structure derived from factor analysis of a measure and impact the reliability and validity of a scale (Clarke, 2000; Hui & Triandis, 1989). Due to the methodological implications and potential influences on the equivalence of an instrument, ERS and response styles are important consideration in cross-cultural research.

Despite the limitations of rating scales including sources of error, they are widely used in cross-cultural research and practice. Given the rapid increase of cultural diversity within the United States, training and research focused on methodological procedures to test the equivalence and validity of rating scales among cross-cultural groups are warranted. Techniques derived from classical test theory and modern test theory are utilized in testing the equivalence of constructs and dimensions of measures across groups (Byrne et al., 2009). The ever-growing immigrant population in the United States

provides a special opportunity for researchers. Although comparing immigrant psychometric data to psychometric data from host countries provides assessment insight, the future of multicultural assessment research may focus on populations from specific geographical areas and the unique characteristics associated with the mixing of cultures in specific regions.

Dimensionality in Measurement

Construct validity is the degree to which a test measures what is intended (Furr, 2018). It relates to the question, "What constructs account for the variation in test performance?" (Cronbach & Meehl, 1955). Construct validity is one of the main types of validity central to test development and encompasses several subtypes of validity. In addition to the subtypes of construct validity, such as convergent and discriminant validity, dimensionality is an assessment of the structural aspect of construct validity (Gessaroli & de Champlain, 2005). In psychological measurement, dimensionality is measured using analytic techniques that evaluate the number of dimensions, or factors, that are estimated by the test's items (Furr, 2018).

Dimensionality is the extent to which an attribute underlies a set of items of a scale (Cronbach & Meehl, 1955; Gessaroli & de Champlain, 2005). The underlying attribute is considered a latent variable, sometimes referred to as a factor, because it is likely an unobserved variable inferred through the measurement of other observed variables, such as test items. When all the items of a scale are assumed to be indicators of the same latent variable, the scale is one-dimensional. One-dimensional scales rely on a composite score of item responses as a measure of the underlying variable. Alternatively, a scale is multi-dimensional when specific items are indicators of different attributes.

Multi-dimensional scales often have an underlying construct such as a higher order variable that is represented by multiple factors indicated by specific items. This is regularly managed by developing subscales that represent each individual factor identified. Typically, these factors, or subscales, come together to represent the general construct of the measure. An example of a widely used multi-dimensional measure can be found in the Parenting Stress Index, Short Form (PSI-SF; Abidin, 2012). The PSI-SF is a multi-dimensional measure of parenting stress commonly used in clinical and research settings. The PSI-SF is a 36-item self-report measure of parenting stress adapted from the 120-item Parenting Stress Index (PSI). The PSI-SF was developed using factor analysis of the PSI, which showed a three-factor solution. Therefore, the PSI-SF includes three subscales that represent three dimensions, or factors, of parenting stress: parental distress, parent-child dysfunctional interaction, and difficult child dimensions. The subscale scores of the PSI-SF provide information related to the source of parental stress going beyond what would be afforded by a total stress score indicating overall severity of stress.

The development and structure of both one- and multi-dimensional measures are often rooted in theoretical models (Abidin, 1992). Measures also aid in assessing the conceptual components of models and are useful in testing theories across groups. Regardless of whether a measure was developed to test a theoretical model or purely for clinical purposes, validation is an essential phase of test construction. Test validation procedures routinely include analyses of reliability and validity. Dimensionality is assessed in order to ensure that the items comprising a test are true measures of the intended attribute across groups (Gessaroli & de Champlain, 2005; Hattie, 1985). However, validation studies occasionally show inconsistent results related to

dimensionality.

There several reasons why discrepancies related to the dimensionality and factor structure of a scale are relevant. Notably, some measurement theorists hold that a composite score that provides an estimate of a corresponding construct is meaningful if the measure or scale has been found to have a one-dimensional structure (Gerbring & Anderson, 1988; Hattie, 1985). Consequently, if a composite score is being relied upon to estimate a certain construct or attribute, uncertainty of the dimensionality of the attribute can lead to errors in measurement and to erroneous conclusions.

Additionally, misinterpreting the dimensionality of a measure can lead to errors in both research and clinical settings. For example, in research settings if responses to scale items are used for group assignment, it must be certain that a composite score is a complete measure of an intended attribute in order to ensure that participants are grouped appropriately. In clinical settings, mistaken assumptions relating to dimensionality and the underlying constructs of a test could have implications for the therapeutic process because scores are often used for screening purposes as well as to inform diagnostic formulation, intervention planning, and treatment monitoring. Further, inaccurate assumptions about dimensionality can lead to over- or under-referral for mental health services and may influence conclusions related to treatment efficacy.

Measuring dimensionality. Two commonly confused concepts related to the validity and reliability of a measure are dimensionality and internal consistency. Cronbach (1951) made distinctions between dimensionality and internal consistency and noted that a test can be interpretable even if the items are not factorially similar. Internal consistency relates to the item homogeneity of a test and how well they combine to

measure a single construct (Davenport, Davison, Liou, & Love, 2015; Henson, 2001). This is not to be confused with the homogeneity of a measure, which references the dimensionality of test. High internal consistency values are not necessarily an indication of unidimensionality, but rather suggest that the items are correlated. In multi-dimensional scales, internal consistency may be high if there is a general factor that underlies the test items. For example, multi-dimensional measures such as the PSI-SF (Abidin, 2012) can have high internal consistency values. The three factors of the PSI-SF are distinct dimensions of parenting stress; however, they correlate with the general construct of parenting stress, which is likely responsible for the reported internal consistency values.

As the terms internal consistency and dimensionality are often inappropriately interchanged, it is not surprising that test statistics of reliability are often drawn into discussions of dimensionality. Coefficient alpha (Cronbach, 1951) is the most commonly used index for reporting reliability (Davenport et al., 2015; Hogan, Benjamin, & Brezinski, 2000). However, theorists argue that alpha is often inappropriately used to assess dimensionality (Davenport et al., 2015; Schmitt, 1996). Specifically, coefficient alpha cannot accurately measure dimensionality due to the possibility of a higher order construct and correlations among common factors that would yield a high alpha. Additionally, there are other considerations when interpreting alpha, such as test length. Davison, Liou, and Love (2015) posited that alpha is not a pure measure of internal consistency because it is also influenced by test length. Therefore, in order to assess the dimensionality of a measure, specific analyses are necessary. There are several available methods to assess dimensionality that generally fall within the categories of Classical

(CTT) and Modern Test Theory.

Classical Test Theory. Factor analysis is a Classical Test Theory (CTT) method often used in test construction and development (Costello & Osborne, 2005). Factor analysis is intended to reveal the underlying factor structure of a group of items while accounting for error and unique variance. Although there are multiple techniques within the realm of factor analyses that can be useful when evaluating the dimensionality of a measure, two of these techniques are frequently used in measurement research. Exploratory factor analytic (EFA) techniques are often used in the early stages of scale development to help identify and to separate dimensions representing theoretical constructs within a domain (Floyd & Widman, 1995). Confirmatory factor analytic (CFA) techniques are used to confirm a theoretical and/or previously derived empirical model (Furr, 2018). CFAs often include theory-based assumptions or findings from previous research and test the assumptions in an effort to confirm a particular factor structure or reveal unexpected factors. EFAs and CFAs are regularly utilized to explore factor structures or to confirm previous findings in scale development and evaluation (Burns & Patterson, 1991; Burns & Patterson, 2000; Furr, 2018; Hu & Bentler, 1999; Weis et al., 2005).

A common misconception is that Principal Components Analysis (PCA) is a form of EFA (Byrne, 2005). Based on the work by Fabrigar, Wegner, MacCallum, and Strahan (1999) as well as Preacher and MacCallum (2003), Byrne (2005) highlighted three conceptual differences between the two. First, the overarching goal of EFA focuses on structural exploration, while the primary goal of PCA is data reduction. In order to explain the pattern of covariance and latent construct(s) underlying a set of variables,

EFA is recommended. Contrastingly, PCA is recommended when it is necessary to reduce a large set of variables to a smaller set of composite variables while maximizing the amount of variance accounted for by the original variables. Variable reduction may be necessary to eliminate collinearity, to simplify data, or to obtain a meaningful summary of the data (Byrne, 2005).

Second, Byrne (2005) emphasized EFA that is a common factor model in which each variable is separated into common variance and unique variance. Unique variance is further conceptualized as including two components – a component specific to that unique variable as well as an error component. PCA neglects to assess unique variance separately and defines each variable as a principal component consisting of both common and unique variance. As principal components represent both common and unique variance, it is inappropriate to view them as representative of latent variables. The ability of EFA to differentiate common variance from unique variance allows researchers to make conclusions relating to the factor structure of datasets (Byrne, 2005).

Lastly, EFA ideally yields a testable model (Byrne, 2005). The identified common factor(s) allow(s) researchers to develop models describing the data and then to test how closely the data fit the model using goodness of fit indices. The PCA does not allow for testing model fit. Therefore, Byrne (2005) concluded that if the goal is to retain linear composites that contain as much shared variance as possible, then PCA is in order. If the goal is to determine interpretable constructs that explain covariance among variables, then EFA is the preferred procedure (Byrne, 2005).

Another consideration in the application of exploratory factor analytic techniques is the use of rotation methods (Byrne, 2005). Orthogonal rotation constrains factors to be

uncorrelated, while oblique rotation allows for correlations among factors. Byrne (2005) argued that although orthogonal rotations yield simpler models, there is more to lose by incorrectly applying an orthogonal rotation compared to an oblique rotation. Incorrectly constraining variables to be uncorrelated can result in misleading estimates. However, utilizing an oblique rotation on truly orthogonal data will still detect independent factors. In addition, many psychological constructs are considered to be correlated in some way, further supporting the use of oblique rotations.

In measurement research and test development, EFA can be followed by CFA. As previously noted, EFA can yield testable models (Byrne, 2005). CFA is not only used to evaluate the overall fit of data to a pre-determined factor model, but also is used to examine a test's internal consistency and reliability (Byrne, 2005; Furr, 2018; Garver & Mentzer, 1999). In order to evaluate model fit, a series of fit indices are considered in CFA (Brown, 2015; Furr, 2018; Hu & Bentler, 1999). Generally, fit indices can be categorized as absolute, adjusted for model parsimony, or comparative or incremental (Brown, 2015). The chi-squared goodness of fit statistic, an absolute fit index, reflects the extent of discrepancy between the actual sample and the covariance of the model being tested (Hu & Bentler, 1999). However, the chi-squared statistic is influenced by sample size. Therefore, additional indices are recommended when evaluating model fit. The standardized root mean squared residual (SRMR; absolute fit), the root mean square error of approximation (RMSEA; parsimony corrected fit), and comparative fit index (CFI; comparative or incremental fit) can be used to evaluate model fit and to avoid the problems of over-relying on the chi-squared fit statistic (Brown; 2015; Furr, 2018). The suggested interpretations for estimating model fit for each statistic recommended by Hu

and Bentler (1999) are shown in Table 2. It should be noted that the interpretations offered by Hu and Bentler (1999) are general guidelines related to fit indices and are not definitive cutoffs (Brown, 2015; Hu & Bentler, 1999).

Table 2

Thresholds for Model Fit Indices in CFA

Statistic	Threshold
SRMR	≤ 0.08
CFI	≥ 0.95
RMSEA	≤ 0.06

Note: Information is based on "Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives" by L.T. Hu and P.M. Bentler, 1999, *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55.

Item Response Theory. Item Response Theory (IRT) is an alternative to CTT for test development and evaluation (Cappelleri, Lundy, & Hays, 2015; Furr, 2018; Kean & Reilly, 2014; Rasch, 1960). In IRT, an individual's response to an item is explained by the respondent's trait level and by qualities of the item (i.e., parameters), such as item difficulty (Furr, 2018; Thomas, 2011). IRT includes a group of measurement models that can increase in complexity as additional item parameters are added, such as item discrimination and guessing parameters (Furr, 2018). In test development and evaluation, IRT can provide valuable information related to the individual, items, and scale that is argued to go beyond that produced by CTT.

In IRT, item parameters include item difficulty, item discrimination, and guessing (Furr, 2018). An item's difficulty is the trait level required to endorse an item. Trait level and item difficulty are both have a mean of 0 and a standard deviation of 1. Item discrimination is the ability of the item to differentiate among respondents based on their trait level (Furr, 2018). Item discrimination values are similar to item-total correlations in CTT, and large, positive discrimination values are favored. Lastly, guessing is the

probability that a person endorses an item purely based on chance (Furr, 2018). Guessing is mostly relevant when items are scored as correct or incorrect.

The differences between CTT and IRT estimations of reliability are important ways in which CTT and IRT vary (Furr, 2018). IRT does not rely on a single measure of reliability as is common in CTT (e.g., coefficient alpha). Rather, in IRT a test might provide better information for some trait levels compared to others. Item information values reflect the probability that a respondent will endorse an item correctly at a particular trait level and can be used to estimate the psychometric quality of an item across varying trait levels. Further, by computing item information values at many trait levels, item information curves (ICCs) can be graphed. These curves can be used to evaluate the psychometric quality of the item and the trait level at which the item provides the most information (Embreston & Reise, 2000; Furr, 2018). In order to understand the psychometric quality of a test as a whole, test information values across varying trait levels can be computed and combined to generate test information curves (Furr, 2018). Test information curves illustrate how a test's psychometric quality can vary across trait levels (Furr, 2018). Ideally, a test would be able to provide good information across varying trait levels.

An illustration of the use of ICCs in supplementing CTT-based conclusions of psychometric properties can be found in Zaidman-Zait et al.'s (2010) item response theory analysis of the Parenting Stress Index-Short Form (PSI-SF; Abidin, 2012). The authors cited the dearth of item-level analyses and homogenous samples in existing research as primary limitations that warranted further psychometric investigation of the PSI-SF (Zaidman-Zait et al., 2010). Specifically, the researchers evaluated the

functioning of the PSI-SF items (i.e., item difficulty and discrimination) across varying levels of parenting stress in a sample of parents of children with Autism Spectrum Disorder (ASD). They employed a non-parametric model which is appropriate for polytomous items and smaller sample sizes. The ICCs showed that the parent distress (PD) subscale items of the PSI-SF functioned well and were useful for assessing the severity of distress among parents of children with ASD at varying levels of stress. However, items in the parent-child dysfunctional interaction (PCDI) and difficult child (DC) subscales functioned less well at discriminating parents across a range of total stress severity within this population. Zaidman-Zait and colleagues (2010) concluded that differences between the study sample (e.g., parents of children with ASD) and the normative sample (e.g., parents of typically developing children) largely explained the differences in scale functioning. Furthermore, they called for caution when using the PSI-SF in atypical populations and for additional research into the content validity of the PSI-SF (Zaidman-Zait et al., 2010). Noteworthy is that the item-level information allowed the authors to understand better how population characteristics can affect the meaning, functioning, and validity of items, and is an example of how IRT can provide information beyond that provided by CTT.

There are a number of IRT models that are used in scale development to evaluate the psychometric functioning of scales (Thomas, 2011). The Rasch model (Rasch, 1960) is a one-parameter model (1PL) and assumes that all items discriminate equally. It is considered to be one of the simplest IRT models and responses are determined by an individual's trait level and one item parameter, the item's difficulty (Furr, 2018; Thomas, 2011; Verhelst & Glas, 1995). Two-parameter models (2PLs) allow for unique item

discrimination and are argued to be more appropriate in clinical assessment and in the measurement of psychological symptoms; however, their use is dependent on large sample sizes which are often unavailable in clinical research (Furr, 2018). Three-parameter models (3PLs) add another parameter referred to as the pseudo-guessing parameter, which accounts for potential guessing.

Rasch models are popular choices for evaluating content specific subscales and testing measurement assumptions for several reasons (Belvedere & Morton, 2010; Thomas, 2011). First, compared to 2PL models, Rasch models require smaller datasets, making them popular choices in clinical research. Second, Rasch models are useful for testing the dimensionality of scale items and investigating item functioning (Tennant, McKenna, & Hagell, 2004). Rasch models assume that all the items of a scale measure a single underlying construct, suggesting unidimensionality, and that the items are *locally independent (*i.e., no correlation among the residuals of the items once the latent variable is controlled for). However, updates to Rasch analysis software allow for further explorations of dimensionality and detection of additional factors (Tennant & Pallant, 2006). Third, Rasch modeling allows for item functioning analysis. In Rasch analysis the items of a scale are presumed to maintain their properties (e.g., item difficulty) regardless of group membership. This assumption can be tested using Differential Item Functioning (DIF) analysis. The assessment of DIF provides information relating to the measurement equivalence of a test's items across groups, such as cultural units or gender and will be further explained below (Furr, 2018; Tennant, McKenna, & Hagell, 2004; Tennant et al. 2004).

A 1PL, or a Rasch model, is appropriate for use with binary response items (Furr,

2018). For polytomous items the Partial Credit Model (PCM; Masters, 1982) or Andrich Rating Scale model (RSM; Andrich, 1978), which belong to the Rasch model family, are appropriate and can be used to assess the rating scale functioning of a measure. Similar to a Rasch model, in a PCM or RSM a person's response to an item is determined by that individual's trait level and the item's difficulty.

There are a variety of global goodness-of-fit statistics to test for model fit in Rasch modeling (Maydeu-Olivares & Montano, 2013; Suarez-Falcon & Glas, 2003). However, a consensus related to the use of one goodness-of-fit statistic over another has not been established, and each statistic has limitations. For example, the R1 and R2 test statistics are specific to Rasch and one-parameter models (Glas, 1988). Both have been found to be more powerful than Pearson's chi-squared statistic to distinguish one-dimensional data from multi-dimensional data (Maydeu-Olivares & Montano, 2013). However, these tests are sensitive to sample size and may not be appropriate to assess unidimensionality in smaller samples. Instead, using multiple indices of fit, such as item fit statistics, PCA of residuals, and detecting differential item functioning (DIF) is an alternative to using global goodness-of-fit tests.

Model fit. Item fit statistics provide two types of information (Smith, Schumacker, & Busch, 1995). First, they estimate misfit for items that an individual is expected to affirm or to deny given their standing on the latent trait. Second, they provide a measure of how susceptible that item is to response patterns inconsistent with the measurement model. There are several item fit statistics that can be used to describe the fit of items to a Rasch model (Smith, Rush, Fallowfield, Velikova & Sharpe, 2008). For example, patterns in item level residuals can be used to test for violations of unidimensionality in

Rasch modeling (Pallant & Tennant, 2007).

Additional fit statistics include the infit mean square (a weighted mean square) and the outfit mean square (an unweighted mean square; Linacre, 2017). Infit is influenced by response patterns, and high infit mean-square values present a threat to internal validity (Linacre, 2017; Smith, 1991). Outfit is most likely influenced by unexpected responses, and high outfit mean-square values are less of a threat to measurement. Both statistics have an accepted range of fit of 0.5 to two, with an expected value of one (Linacre, 2017). Mean-squares greater than one suggest that there is "more variation" in the data than predicted by the model and underfit (Linacre, 2017). Mean-squares less than one indicate that the data may overfit the model (Linacre, 2017).

The fit statistics generated from Rasch modeling techniques are vulnerable to variations in sample properties (Linacre, 2017). Specifically, mean-square statistics will move closer to the expected number, one, as sample size increases, which may cause misinterpretations about fit. In order to account for sample size, the infit and oufit mean-square can be converted to standardized *t*-statistics using the Wilson-Hilfrey transformation (Bond & Fox, 2015; Linacre, 2017). Standardization of the mean-square statistics takes sample size into account by including the mean and variance of the mean-square value. These statistics have an acceptable range between plus or minus two, with an expected value of zero. Values closer to or greater than two demonstrate more variance than predicted and values closer to or less than negative two demonstrate highly constrained items (Bond & Fox, 2015).

When making conclusions relating to the dimensionality of the data using fit statistics, the unique limitations of the values must be considered. Tennant and Pallant

(2006) found that Rasch model fit statistics did not identify misfitting items in the presence of two interrelated factors. Yu, Popp, DiGangi, and Jannasch-Pennell (2007) found additional evidence that questioned the use of residual cut-off scores for assessing unidimensionality in the presence of correlated factors. In addition, infit and oufit statistics are susceptible to "accidents" in the data, such as guessing. Due to these limitations, assessing unidimensionality using a variety of indices is recommended (Linacre, 2017).

Residual-based PCA. In addition to utilizing fit indices, residual-based PCA can be used to assess dimensionality (Linacre, 2017; Wright, 1996). This process involves looking for patterns in data that may not adhere to the Rasch assumptions. Contrary to the goal of common factor analysis, residual-based PCA does not aim to construct variables but rather to explain variance. The single latent trait in Rasch modeling is considered to be the *Rasch Factor*. By comparing the ratio of variance explained by the Rasch Factor to that explained by the residual factors, the possibility of a second underlying construct can be tested (Allison, Baron-Cohen, Wheelwright, Stone, & Muncer, 2011; Linacre, 2017). Although a non-traditional approach, Rasch analysis followed by PCA of residuals has been found to be more effective at identifying multi-dimensionality when compared to factor analysis of response-level data alone (Linacre, 2017).

*DIF***.** IRT provides item functioning information useful to understanding overall test score differences in cross-cultural assessment (Cauffman & MacIntosh, 2006; Tennant et al., 2004). In order to make meaningful comparisons of scores between groups, the structure of the measure and the functioning of the items must be assumed to be invariant. The presence of DIF suggests that an item's properties in one group (e.g.,

gender, socioeconomic status, or ethnicity) are different from the item's properties in another group, and, as a result, the probability of endorsing the item varies based on group membership (Furr, 2018). This is problematic when making comparisons across groups and is a threat to measurement equivalence (Tennant, McKenna, & Hagell, 2004). Overall, the presence of DIF between groups on a particular item indicates that meaningful comparisons cannot be made between responses to that particular item (Furr, 2018).

Significant DIF can indicate a violation of the assumption of unidimensionality (Ackerman, 1992). If a scale is not one-dimensional enough, nuisance factors can influence the measurement of the latent trait. In the presence of DIF, other factors may be driving response patterns alternative to the latent trait, implying that a scale is not measuring the same trait for all respondents (Walker, 2011). By comparing the item parameters between groups, conclusions can be made about the functioning of the item in relation to group membership. Ideally, group membership does not significantly influence item difficulty. If the DIF is substantial, score comparisons between groups may not accurately reflect estimates of the latent variable.

In the presence of DIF, items can be removed if the impact on the overall test score is significant (Langer, Hill, Thissen, Burwinkle, Varni, & DeWalt, 2008). This approach to addressing DIF is problematic for scales with few items. Therefore, assessing the impact of the item on the scoring of the scale is an alternative approach to managing the presence of DIF while maintaining scale length (Linacre, 2017). Investigating for possible cancellation effects (e.g., the differential item functioning for one item, is cancelled out by the differential item functioning of another item) or rewording an item in

order to eliminate DIF are also alternative approaches to managing the presence of DIF while maintaining the items in a scale (Linacre, 2017).

Rasch modeling examples. Supplementing classical test theory techniques with Rasch modeling is an emerging trend in measurement research (Thomas, 2011). Rasch modeling provides an alternative analytic approach when more commonly used measurement analysis techniques have been exhausted. Historically, the focus of graduate training in measurement analysis has largely been on the use of CTT to investigate test dimensionality, and researchers may be reluctant to endeavor to learn alternative test theories. However, plenty of published studies which use Rasch modeling are available for review (e.g., Wardenaar et al., 2010). The examples found in the literature provide a potential framework for researchers considering the application of IRT and demonstrate the utility of modern test theory applications.

Wardenaar and colleagues (2010) utilized Rasch modeling in addition to CFA to supplement their understanding of the factor structure and dimensionality of a measure. Similarly, Cauffman and MacIntosh (2006) demonstrated the use of Rasch modeling to investigate the cross-cultural use of an instrument. Specifically, these two studies emphasized the utility of using Rasch modeling to supplement existing research. Further, Rasch modeling is showcased in the studies as an alternative analytic technique that can be useful when previous investigations have produced mixed results (Cauffman &MacIntosh, 2006; Wardenaar et al., 2010).

The Inventory of Depressive Symptomatology Self Report (IDS-SR; Rush, Gullion, Basco, Jarrett, & Trivedi, 1996) was developed as a one-dimensional measure of depression. However, Wardenaar and colleagues (2010) found inconsistent results

relating to the factor structure of the IDS-SR in research literature. Therefore, in an effort to find a stable factor model for this instrument CTT methods and Rasch modeling were utilized. Specifically, the researchers aimed to find the best-fitting factor model for the data and to assess the dimensionality of the IDS-SR.

The CTT analysis of the IDS-SR data set included a CFA of a one-, two-, three-, and four-factor model proposed by previous investigations across four diagnostic groups (Wardenaar et al., 2010). The three-factor solution was found to provide the best fit to the data, as indicated by the indices-of-fit across groups. The Rasch analysis included a total scale analysis, which indicated that 10 of the 28 items poorly fit the model. Then, each of the three factors from the CFA were individually fit to the Rasch model to investigate whether those factors could be used as subscales. As part of the analysis, items with poor fit were eliminated in order to improve the unidimensionality of the potential subscales. In addition, DIF was assessed to evaluate the generalizability of item functioning across groups. This was followed by PCA of the residuals to explore the unidimensionality of each proposed scale.

Wardenaar et al. (2010) concluded that the IDS-SR functioned best as a multidimensional measure of depression with only two unidimensional subscales. The analytic strategy employed by Wardenaar and colleagues demonstrated the utility of Rasch modeling in scale development. Although PCA of the IDS-SR data, along with CFA, found that a three-factor solution best fit the data, Rasch analysis of the IDS-SR data and item functioning resulted in two one-dimensional independent subscales. Had the researchers relied solely on CTT techniques, it is likely that the third factor that was ultimately dropped due to poor fit to the model would have been retained.

In another study demonstrating the utility of Rasch modeling, Cauffman and MacIntosh (2006) investigated the cross-cultural use of a juvenile screening instrument using Rasch analysis. Item fit and DIF statistics were used to evaluate the items of the seven subscales of the Massachusetts Youth Screening Instrument, second version (MAYSI-2; Grisso, Barnum, Fletcher, Cauffman, & Peuschold, 2001) within an ethnically diverse sample. The authors found several of the subscales contained misfitting items. Additionally, clinically significant DIF was found across gender and ethnic groups. Cauffman and MacIntosh concluded that these deviations from the Rasch model were evidence of multi-dimensionality and determined that the MAYSI-2 subscales may not be entirely one-dimensional. Moreover, a number of the items performed differently based on the respondent's ethnicity, suggesting a lack of measurement equivalence across groups.

Several notable conclusions from the Cauffman and MacIntosh (2006) findings can be made. First, several of the items on the MAYSI-2 demonstrated significant misfit and no identified DIF, indicating that misfit and DIF can exist independent of one another. Second, meaningful comparisons between scores are problematic, because the presence of DIF suggests that the properties of the items vary across groups. Lastly, the findings highlight the importance of rigorous exploration of the psychometric properties of a measure that is to be used among diverse populations. Similar to the majority of screening measures, the MAYSI-2 was originally normed using predominately non-Hispanic white youths and has been found to have good psychometric properties using CTT analysis (Grisso, Barnum, Fletcher, Cauffman, & Peuschold, 2001). However, findings from Cauffman and MacIntosh's (2006) Rasch analysis suggest that the

MAYSI-2 may perform differently when used with diverse populations, regardless of established psychometric properties.

In summary, CFA and Rasch modeling are two methods used in test development and test evaluation. Each has its respective strengths and limitations. With regard to application, CFA, and other CTT techniques are more likely to be taught in training settings compared to Rasch modeling and other IRT methods (Thomas, 2011). Therefore, CFA is more often utilized, in part due to the level of familiarity with this analytic method. Further, some IRT methods require significantly larger sample sizes, and datasets must meet more rigorous assumptions, such as unidimensionality, limiting the applicability of IRT to some samples (Cappelleri et al., 2015).

By comparison, CFA is sample-dependent, suggesting that the findings from one sample may not hold for different samples (Abrahamse et al., 2015). Rasch modeling can provide information beyond what can be gained from CFA by way of reliability and individual item functioning across trait levels; however, many helpful measures currently in use have been developed using CTT methods. So, utilizing techniques from both traditions to complement each other based on their strengths may be the best approach to analytic strategies (Cappelleri et al., 2015; Kean & Reilly, 2014).

The Eyberg Child Behavior Inventory

The Eyberg Child Behavior Inventory (ECBI) is a 36-item parent rating scale used to assess disruptive behavior problems in both children and adolescents (Eyberg & Robinson, 1983). The ECBI includes a list of 36 typically occurring problem behaviors reported by parents of children with conduct problems (Eyberg & Ross, 1978). The ECBI is designed to identify children and adolescents with conduct problems and is commonly

used to monitor treatment effects. The inventory helps assess behaviors on two dimensions, the frequency of occurrence, i.e., the Intensity Scale, and the identification of the behavior as a problem by the reporter, i.e., the Problem Scale. For both scales, Eyberg and Pincus (1999) provided clinical cutoff points which indicate when further evaluation of problematic behaviors is warranted. Parent responses greater than a raw Intensity Scale score of 131 or a Problem Scale score of 15 are considered to be in the clinical range for disruptive behavior problems (Eyberg & Pincus, 1999).

The ECBI was originally created to meet the need of therapists treating children with behavior disorders (Eyberg & Robinson, 1983; Eyberg & Ross, 1978). At the time of development, there was a demand for a way to assess problem behaviors commonly reported by parents of conduct-disordered children. Early validation studies of the ECBI found that this instrument is capable of differentiating between conduct problem children and typical children (Eyberg & Ross, 1978). In addition to the fact that the ECBI filled a need, its brevity and scoring ease added to its positive reception in the field of child assessment. Today, the ECBI is widely used for assessment, intervention, and research purposes in a variety of treatment settings (Berkovits, O'Brien, Carter, & Eyberg, 2010; Funderburk et al. 2003). The properties of the ECBI are summarized in Table 3.

Table 3

Properties of the ECBI

Ages	Items and Scoring	Cutoff Scores	Reliability and Validity	Languages
2-16	36 items scored on a 7-point scale (Intensity); scored again on a Yes/No scale (Problem)	131 (Intensity) 15 (Problem)	High on Internal Consistency, Test-retest Reliability, Convergent Validity, and Discriminative Validity	Chinese English German Japanese Korean Lebanese Norwegian Russian Spanish Swedish

Note: ECBI = Eyberg Child Behavior Inventory. Information is based on published information in the ECBI professional manual by Eyberg & Pincus, 1999 and restandardization study by Colvin, Eyberg, & Adams, 1999.

The Spanish version of the Eyberg Child Behavior Inventory. A Spanish version of the ECBI was developed using translation and back translation methods (Garcia-Tornel et al., 1998; Gross et al., 2007). It is commercially available for purchase through the publisher. There is comparatively much less research related to the psychometric properties of the Spanish version than the English version of the ECBI. Reliability and validity studies have been completed using Spanish samples (Fernández, Gorostiza, Lafuente, Ojembarrena, & Olaskoaga, 1998; Garcia-Tornel et al., 1998). However, the information relating to the validity of the Spanish ECBI for Hispanic-Americans is sparse. The extant research includes two studies primarily investigating the functioning of the Spanish version of the ECBI in a sample drawn from Spain, and one study focused on the psychometric properties of the ECBI in a Hispanic-American sample (Fernández, Gorostiza, Lafuente, Ojembarrena, & Olaskoaga, 1998; Garcia-Tornel et al., 1998; Gross et al., 2007).

The studies including Spanish samples show that the ECBI is a useful tool in screening for disruptive behaviors among Spanish children (Fernandez et al., 1998; Garcia-Tornel et al., 1998). The findings demonstrate that the Spanish version of the ECBI has good internal consistency and test-retest reliability (Fernandez et al., 1998; Garcia-Tornel et al., 1998). Relating to dimensionality, results from a PCA by Fernández et al. (1998) showed four components that accounted for 84 percent of the variance collectively, and the first component accounted for 49 percent of the variance.

Overall, these findings suggest that the Spanish version of the ECBI is a reliable measure of disruptive behaviors in Spanish populations. Still, the findings suggest the possibility of a multi-dimensional structure. These results also provide necessary information relating to the norms for Spanish populations. Garcia-Tornel and colleagues (1998) proposed preliminary suggestions for modifications to the U.S.-based norms pending additional research. However, in order to recommend the use of Spanish-based norms, additional research replicating these findings is necessary.

In the United States, the Spanish version of the ECBI has been used to measure parent child interactions and Parent Child Interaction Therapy outcomes in Mexican American and Puerto Rican families (Borrego et al., 2006; Matos, 2006; McCabe et al. 2010; McCabe et al., 2012). The Spanish version of the ECBI has demonstrated high internal consistency within Hispanic samples. However, just one investigation of the psychometric properties of the Spanish version of the ECBI with a United States-based sample was found (Gross et al., 2007).

Gross et al. (2007) investigated the reliability and validity of the ECBI in African American and Hispanic families, due to the lack of ethnic diversity in existing

psychometric research of the ECBI. First, Gross et al. used a *t* – test to examine mean differences between the Spanish and English versions of the ECBI among low-income Hispanics. No significant mean differences were noted on the Intensity or the Problem Scale. The investigators concluded that the Spanish version of the ECBI appears to function similarly to the English version in Hispanic samples. However, the authors found scale differences by ethnicity. Specifically, Hispanic parents were more likely than non-Hispanic White parents to score a behavior as a problem when the frequency of the behavior was rated as "never" or "seldom" occurring. Given the scale differences, Gross et al. recommended that additional research exploring the construct validity of the ECBI among Hispanic parents of preschool children was needed in order to form definitive conclusions.

Aside from international samples, there is little research on the psychometric properties of the Spanish version of the ECBI. The norms proposed by Garcia-Tornel et al. (1998) provide Spanish cutoff scores of 124 for the Intensity Scale and 10 for the Problem Scale. However, use of these norms with Hispanic-American samples may be inappropriate due to cultural differences. Moreover, the findings by Gross et al. (2007) suggest that both versions of the ECBI function similarly within Hispanic-American samples. Therefore, there is precedent for generalizing the findings of the English version of the ECBI to the understanding of the Spanish version of the ECBI in Hispanic-American samples.

Standardization of the Eyberg Child Behavior Inventory. The ECBI was originally standardized for children in 1980 and for adolescents in 1983 (Eyberg & Robinson, 1983; Robinson, Eyberg, & Ross, 1980). Problematic behaviors listed in the

ECBI are rated along two dimensions, problem and intensity. The two composite scores of the ECBI reflect two different estimates of the number and type of problems and the intensity of problems relating to child behavior difficulties. Findings have consistently demonstrated good discriminant and concurrent validity, and the ECBI is considered a well-validated measure of child conduct problems (Boggs et al., 1990; Funderburk et al., 2003). Originally, the ECBI was conceptualized as a one-dimensional measure, which indicates that all 36 items of the ECBI assess one underlying construct (Eyberg & Pincus, 1999). However, there is ongoing debate regarding the underlying factor structure of the ECBI.

More recent standardization studies completed by the developers of the ECBI continue to find evidence that supports a one-dimensional construct, and the one factor structure has been replicated in independent investigations (Abrahamse et al., 2010; Butler, 2011; Colvin et al., 1999; Gross et al., 2007). Yet, evidence in support of a multi-dimensional model continues to add to the criticism of the ECBI's construct validity across samples (Burns & Patterson, 1991; Burns & Patterson, 2000; Jeter, Zlomke, Shawler, & Sullivan, 2017; Stern, 2007; Weis et al., 2005). Specifically, the findings by Burns and Patterson (1991, 2000) supporting a three-factor structure (i.e., oppositional defiant behaviors, inattentive symptoms of ADHD, and conduct problem behaviors) and a reduced 22-item measure have garnered the most attention. The three-factor model of the ECBI appears to have sparked research interest and additional investigation of the ECBI's factor structure.

Dimensionality of the Eyberg Child Behavior Inventory. The psychometric properties of the ECBI have been explored in several U.S.-based and international

samples. The majority of the U.S. investigations have included predominately non-Hispanic European American samples with some exceptions. Overall, six investigations of the ECBI's factor structure have found evidence for a univariate factor model (Abrahamse et al., 2015; Butler, 2011; Colvin et al., 1999; Eyberg and Robinson, 1983; Gross et al. 2007; Robinson et al., 1980). In comparison, the results from seven studies showed that the ECBI is a multi-dimensional measure. Specifically, the researchers found support for a three-factor (Axberg et al., 2008; Burns & Patterson, 1991, 2000; Stern, 2007; Weis et al. 2005), four-factor (Jeter et al., 2017), and a bifactor model with three specific factors and one general factor (Hukkelberg, 2017). Axberg et al. (2008) and Weis et al. (2005) found adequate fit of the three-factor model with 22- items proposed by Burns and Patterson (2000) to their data. Additionally, Stern (2007) found that an alternative three-factor model with 25-items demonstrated good fit to the data.

Evidence for a one-dimensional measure. Investigations of the ECBI's factor structure have increased since Burns and Patterson (1991, 2000) published their findings supporting a three-factor model of disruptive behavior problems. In a re-evaluation of the ECBI, Colvin et al. (1999) found that one principal component explained the majority of variance in ECBI scores and re-affirmed the unidimensionality of the ECBI. Similarly, Gross et al. (2007) and Butler (2011) found evidence for a one-factor model of the ECBI in Hispanic, African American, and non-Hispanic samples. In contrast to the previous studies, Abrahamse et al. (2015) utilized Rasch analysis in addition to CFA to confirm the one-factor solution in a Dutch sample.

Based on this research, it is difficult to make definitive conclusions about the ECBI's factor structure because of methodological concerns. Mainly, a number of the

investigators utilized PCA to assess dimensionality, and the majority of samples are predominately drawn from non-Hispanic European American and international populations. When considering the findings using PCA, conclusions made relating to the dimensionality of the latent structure should be interpreted with caution. Specifically, PCA does not differentiate between unique and common variance. Therefore, the findings are not representative of the latent variable but rather of a component. Additionally, aside from the study by Gross et al. (2007), Hispanic individuals made up a small fraction of the research samples.

Classical Test Theory approaches. Initial ECBI normative data come from two studies (Eyberg & Robinson, 1983; Robinson, Eyberg, & Ross, 1980). The earlier study (Robinson, et al., 1980) included a sample of 512 children two to 12 years old who were seen in an outpatient pediatric clinic in the northwestern United States. The later study (Eyberg & Robinson, 1983) involved 102 adolescents 13 to 16 years old who were brought by their parents to an outpatient pediatric clinic located in a northwestern university health sciences center. Both samples included predominately non-Hispanic European American children with a variety of behavior problems, developmental delays, and chronic illnesses. Using PCA methods, Robinson et al. (1980) showed that 68% of the variance in their data was explained by the first factor, and all 36-items of the ECBI loaded positively onto the dominant factor. Similarly, Eyberg and Robinson (1983) found that a first factor accounted for 54% of the variance using a principal components analysis of ECBI data. The results provided initial support for the marketing of the ECBI as a unidimensional measure of conduct problems in children. Notably, the researchers cautioned that conclusions made from the data might not generalize to non-European

American families, as they were underrepresented in the study.

Two additional standardizations of the ECBI were independently completed using more ethnically varied samples (Burns & Patterson, 1990; Burns, Patterson, Nussbaum, & Parker, 1991). Burns and Patterson (1990) found that a principal component accounted for 29.4% of the variance in the Intensity Scale and 24% of the variance was accounted for by a principal component in the Problem Scale. Similarly Burns et al. (1991) found that 30.2% of the variance was accounted for a by the first principal component in the Intensity scale and 24.6% in the Problem scale. These data provide additional evidence supporting the ECBI as a psychometrically sound measure of conduct-problem behaviors in children and adolescents.

Similar to the criticisms of the previous research, the analytic strategy and sample characteristics call into question the generalizability of the findings. In fact, the norms reported by Burns and Patterson (1990) and Burns et al. (1991) have been criticized as not being representative of the population generally studied, since the majority of children included in the samples had no history of treatment for learning disability, behavior problems, or chronic illnesses (Achenbach, 2001; Colvin et al., 1999). In addition, Burns and Patterson (1990) commented on the overrepresentation of non-Hispanic European American children (78% European American) in their total sample, while the sample in Burns and colleagues (1991) investigation was 90% European American. Overall, both samples had overwhelming rates of non-Hispanic European American participants with limited inclusion of African American participants and an even smaller Hispanic presence.

The ECBI underwent a second formal standardization process that included a sample of 798 children ages two to 16 (Colvin et al., 1999). Although this sample was somewhat more ethnically diverse, it included predominately non-Hispanic European American children (i.e., 74% European American). For the entire sample, Cronbach's alpha was 0.95 for the Intensity Scale and 0.93 for the Problem Scale. Results from PCA showed that 33 of the 36 items loaded onto a strong first factor, which demonstrated that the majority of the ECBI items measured a uniform latent variable. Although subject to the same limitations as the previous research, the authors noted that their sample resembled U.S. Census data at the time. Therefore, the updated normative data generated from this study was most likely generalizable at least to families in the Southeastern U.S.

The investigation by Gross and colleagues (2007) was the first study of the ECBI in an ethnically diverse sample. The purpose of this study was to provide additional evidence for the reliability, validity, and factor structure of the ECBI and the Spanish version of the ECBI for children from different ethnic backgrounds. A sample of 682 parents of two- to four-year-olds was recruited from a Chicago metropolitan area. The sample included African Americans (28.7%), Hispanics (46.8%), and non-Hispanic European Americans (24.5%). The Hispanic group was comprised of primarily Mexican American participants, which was consistent with the population of Hispanics in the area. Results showed no significant mean differences between the Spanish and English versions of the ECBI within the Hispanic group; therefore, both groups were combined for factor analysis. Comparisons between CFA results using a one-factor and a three-factor model demonstrated that the Burns and Patterson (2000) three-factor solution, including factors of oppositional defiant behaviors, inattentive symptoms of ADHD, and

conduct problem behaviors, fit significantly worse than the one-factor solution. These results were the first to support the unidimensionality of the ECBI among ethnic minority groups.

More recently, Butler (2011) replicated the one-factor structure using a 25- item ECBI proposed by Stern (2007) in a sample of low-income African American and non-Hispanic white preschoolers. Results from EFA and CFA demonstrated that a single-factor structure best fit the data from both groups. The investigation by Butler appeared to be based on the work by Stern, who found the ECBI to be multi-dimensional. Therefore, Butler concluded that the population-dependent nature of psychometric properties indicated by CTT techniques (Haynes, Nelson, & Blaine, 1999), varied response styles (Van de Vijver & Poortinga, 1997), and perceptions of the underlying construct across groups (Hillemeir, Foster, Heinrichs, & Heier, 2007) might explain the inability to replicate a three-factor structure. Nevertheless, Butler concluded that a three-factor solution is not recommended to screen for specific behavior problems among low income African American and non-Hispanic European American populations.

Item Response Theory approaches. Factor analysis is predominately used in the research relating to the ECBI's dimensionality. However, given the limitations of factor analytic techniques, it is not surprising that some researchers have turned to alternative analysis strategies. As previously noted, item response theory approaches, such as Rasch modeling, provide an alternative analytic strategy for test measurement.

Abrahamse et al. (2015) evaluated the dimensionality of the ECBI's Intensity and Problem Scales using Rasch modeling in a Dutch sample. The study included a community sample (n = 326) and a clinically referred sample (n = 197) of parents of

children ages two to eight years old. The community sample primarily identified as Dutch (90.8%), while the clinically referred sample included participants from a range of ethnic backgrounds, including 43.5% non-Western participants (mainly Moroccan and Turkish). An EFA with an oblique rotation was used to explore the dimensionality of the ECBI. Next, a Rasch model was employed to test the fit of the one-factor model found by the EFA.

The EFA of the ECBI Intensity Scales revealed a dominant first factor explaining 30.7% of the variance in the community sample and 32.1% in the clinical sample (Abrahamse et al., 2015). Nine factors were identified using eigenvalues greater than one, with a sharp dominance observed in the first factor (i.e., 11.2 in the first factor compared to 2.1 in the second factor). The EFA of the Problem scales of the two samples yielded similar results, with a dominant first factor accounting for 30.0% of the variance in the community sample and 25.3% in the clinical sample. Several factors were also identified in both samples, with raw eigenvalues supporting a strong first factor. Scree plots also supported the presence of one dominant factor. Then, Abrahamse et al. (2015) combined the samples for the Rasch analysis in order to increase the sample size. The total sample for the Intensity Scale was $N = 514$ and $N = 481$ for the Problem Scale. Results showed good overall fit to the Rasch model for both scales, and the authors concluded that the ECBI was a unidimensional measure of problematic behaviors in children.

The use of alternative analysis techniques to improve the clarity of the psychometric properties of the ECBI is a strength of the study by Abrahamse and colleagues (2015). However, there are several concerns related to the results reported by the authors. For example, Abrahamse et al. used item-oriented fit statistics, *S*-tests, and

the R1c statistic, an overall fit statistic, to evaluate the dimensionality of the ECBI. The *S*-test and the R1c statistic are first-order statistics (Suarez-Falcon & Glas, 2003). The R1c statistic is derived from chi-square. Therefore, it is vulnerable to the same threats as a chi-square statistic, such as sample size and test length. Abrahamse et al. (2015) did not discuss additional item fit statistics from the Rasch model that may have provided further information regarding item functioning. Further, despite the authors' aim to investigate the cross-cultural use of the ECBI in a predominately Dutch population, DIF was not discussed.

Evidence for a multidimensional measure. In 1991, Burns et al. alluded to the relationship between the 36 items on the ECBI and the diagnostic criteria for externalizing behavior disorders. Specifically, the researchers suggested that the ECBI behavior disorder criteria reflected in the items mirror the criteria listed in the Diagnostic and Statistical Manual of Mental Disorders used at the time of the study (DSM-III-R; American Psychiatric Association, 1987). Therefore, Burns and Patterson (1991) investigated whether three dimensions that are reflective of the DSM categories would emerge from a factor analysis of ECBI data.

Using data collected from 1,526 children in four northwestern states, Burns and Patterson (1991) utilized a PCA with varimax rotation on the intensity scores of the ECBI. Results showed seven components with eigenvalues greater than one. The authors retained three components for rotation based on the goal to test for three conceptually supported dimensions of disruptive behaviors. The factor loadings supported three components characterizing attention difficulties, oppositional defiant behavior, and violation of basic rights of others through overt aggression. Next, in order to replicate

these findings on a random sample, Burn and Patterson collected ECBI data from children in a Seattle school district. Following the same analytic plan, eight components emerged with eigenvalues greater than one. After retaining three components for varimax rotation, a three-factor solution consistent with their previous findings emerged. The findings led the authors to conclude that the ECBI is a multi-dimensional measure, and additional research was recommended to clarify the organization of disruptive behaviors (Burns & Patterson, 1991).

In order to assess further the factor structure of the ECBI, Burns and Patterson (2000) combined two predominately non-Hispanic European American (i.e., 85 %) samples of children and adolescents. The investigators randomly created two sample groups in order to complete an exploratory factor analysis (EFA) with the first group and a confirmatory factor analysis (CFA) with the second group. Following the EFA, a four-factor solution was found. The factors included oppositional defiant behaviors, inattentive symptoms of ADHD, conduct problem behaviors, and a fourth unclear factor. The factor loadings of the 36 items and the four factors from the EFA are summarized in Table 4. In the CFA phase, only the three meaningful factors and the relevant 22 items were included. As hypothesized, the CFA resulted in a reasonable fit of the three-factor model. Burns and Patterson argued that clinicians and researchers could use the identified subscales as a more meaningful measure of behavior problems in screening, outcome measurement, and research procedures.

Table 4

Factor Loadings for Exploratory Factor Analysis with Maximum Likelihood Extraction

Items	F1	F2	F3	F4
F1: Oppositional Defiant Behavior Toward Adults				
11. Argues with parents about rules	**0.92**	-0.02	0.03	-0.25
10.Acts defiant when told to do something	**0.91**	-0.07	-0.03	-0.01
9. Refuses to obey until threatened with punishment	**0.73**	0.02	0.02	0.07
14. Sasses adults	**0.73**	-0.14	0.05	0.03
5. Refuses to do chores when asked	**0.65**	0.00	-0.04	0.01
12. Gets angry when doesn't get own way	**0.63**	-0.04	0.02	0.21
8. Does not obey house rules on own	**0.62**	0.19	0.13	-0.12
7. Refuses to go to bed on time	**0.48**	0.11	-0.13	0.12
13. Has temper tantrums	**0.46**	-0.01	-0.01	**0.36**
17. Yells or screams	**0.39**	-0.12	0.19	**0.30**
6. Slow in getting ready for bed	**0.32**	0.18	-0.11	0.13
3. Has poor table manners	0.22	0.14	0.22	0.09
F2: ADHD Behavior				
31. Has short attention span	-0.06	**0.95**	-0.11	0.03
30. Is easily Distracted	-0.13	**0.90**	-0.03	0.02
34. Has difficulty concentrating on one thing	-0.07	**0.88**	-0.01	-0.04
32. Fails to finish tasks or projects	0.09	**0.71**	-0.01	-0.05
35. Is overactive or restless	0.11	**0.43**	0.10	0.10
33. Has difficulty entertaining himself or herself alone	0.06	**.034**	0.10	0.11
F3. Conduct Problem Behavior				
25. Verbally fights with sisters and brothers	-0.04	-0.17	**0.76**	-0.07
27. Physically fights with sisters and brothers	-0.11	-0.14	**0.75**	0.09
23. Teases or provokes other children	0.03	0.01	**0.72**	-0.13
24. Verbally fights with friends his or her own age	-0.02	-0.02	**0.64**	0.04
26. Physically fights with friends his or her own age	-0.08	0.02	**0.63**	0.10
22. Lies	0.14	0.19	**0.48**	-0.17
21. Steals	0.16	0.11	**0.39**	-0.15
19. Destroys toys and other objects	0.11	0.17	**0.39**	-0.06
20. Is careless with toys and other objects	0.14	0.26	**0.33**	0.01
F4. Unamed Factor				
15. Whines	0.02	-0.06	-0.02	**0.75**
16. Cries Easily	-0.19	-0.06	0.16	**0.69**
2. Dawdles of lingers at mealtime	-0.01	0.07	-0.18	**0.62**
29. Interrupts	0.09	0.20	0.08	**0.36**
1. Dawdles in getting dressed	0.11	0.18	-0.16	**0.36**
18. Hits parents	0.19	-0.07	-0.04	**0.35**
28. Constantly seeks attention	-0.02	0.21	0.26	**0.33**
4. Refuses to eat food presented	0.31	-0.12	-0.08	**0.32**
36. Wets the bed	-0.03	0.10	-0.03	**0.30**
Percentage of Variance	32.59	6.85	5.75	4.33

Note: N=1,263 children. ECBI=Eyberg Child Behavior Inventory. ADHD=Attention-Deficit/Hyperactivity Disorder. F1=Factor 1; F2=Factor 2; F3=Factor 3. Factor loadings greater than 0.29 are shown in bold. Adapted with permission from "Factor structure of the Eyberg Child Behavior Inventory: Unidimensional or multidimensional measure of disruptive behavior" by G. Burns and D. Patterson, 2000, *Journal of Clinical Child and Adolescent Psychology*, *20*(4), 439-444

Based on the findings by Burns and Patterson (1991; 2000), Weis et al. (2005) investigated the proposed three-factor structure of the ECBI in a clinical sample. The researchers hypothesized that in their sample of parents of young children, a one-factor model would best fit the ECBI data. Their sample was primarily non-Hispanic European American (i.e., 85%) with 10% African American and 2% Hispanic participants residing in Midwestern cities and rural areas. Although a one- and a two-factor structure fit the data, ultimately the three-factor structure previously identified by Burns and Patterson (2000) was found to have the strongest support of model fit.

Weis and colleagues (2005) not only found support for the tripartite model proposed by Burns and Patterson (1991; 2000), but their investigation also provided the initial evidence for the clinical utility of the three-factor model and evidence for the discriminant validity of the three factors identified (Burns & Patterson, 1991, 2000; Weis et al. 2005). Overall, all three factors were found to have adequate negative predictive power, and the Inattentive and Oppositional factors were found to have adequate positive predictive power in identifying children with behavior problems from a normal population (Weis, et al. 2005). However, the factors were not effective in differentiating children's behaviors within the externalizing spectrum. The results found by Weis et al. (2005) show that an alternative interpretation of ECBI data could have implications for both research and clinical settings.

The most recent study proposing a three-factor model for the ECBI is an unpublished exploratory factor analysis of data from parents of 181 children referred to an ADHD clinic in northern Florida (Stern, 2007). Similar to previous research, the sample included 75% European American, 18% African American, 2% Hispanic

participants and 5% “other” or “unspecified.” Results showed that a three-factor model best fit the data, and 25 items explained the majority of the variance. The factors found represented oppositional defiant behaviors, attention difficulties, and conduct problems. Stern (2007) went further in her investigation and found that the three factors demonstrated good internal consistency and strong evidence for convergent validity.

In general, the four investigations of the ECBI that found support for multi-dimensionality demonstrate similar methodological concerns. Although the sample in Stern’s (2007) study included a higher percentage of African American participants and was most likely representative of the area of data collection, Hispanic participants were underrepresented across all the samples. Specifically, the generalizability of the findings across groups is questionable due to the limited diversity in the samples. Additionally, the analytic procedures are vulnerable to the same threats as those used in the majority of the research that found support for unidimensionality.

Bifactor model. Aside from the Abrahamse et al. (2015) study, the majority of authors evaluating the factor structure of the ECBI have focused on traditional factor analytic strategies such as CFA. Hukkelberg (2017) noted that relying on traditional CFAs to evaluate the factor structure of the ECBI does not bridge the gap between the discrepant findings supporting a unitary construct and a tripartite model of the ECBI data. In order to address this concern, Hukkelberg hypothesized that a bifactor model, which allows for the differentiation of shared and unique variance among items, may better categorize disruptive behaviors as measured by the ECBI.

Referred to as a latent bifactor approach, the bifactor model can be used to determine to what extent disruptive behavior problems are better understood as a general

construct or a latent construct representing common variance across oppositional defiant, inattentive, and conduct behavior (Hukkelberg, 2017; Reise, 2012). In order to explain this conundrum, Hukkelberg (2017) considered three factor models, a traditional three-factor CFA, a bifactor CFA, and a bifactor exploratory structural equation model (ESEM) using Burns and Patterson's (2000) tripartite solution with 22- items. The ESEM model was included to assess more accurately the construct validity of the ECBI by comparing the goodness of fit between models that restrict cross-item loadings (bifactor CFA) and one that allows for it (bifactor ESEM). The sample consisted of 353 children enrolled in either a brief parent training (BPT; $n = 137$) or the Oregon model of parent management training (PMTO; $n = 216$) intervention recruited from five health regions in Norway. Both samples included children three to 12 years of age with parents of primarily Norwegian background, and the PMTO sample included children with higher levels of problem behaviors compared to the BPT group.

Hukkelberg (2017) found the traditional three-factor CFA to fit poorly in both the BPT and the total sample. When compared to the bifactor ESEM, the bifactor CFA model, with a general problem behavior factor and three specific factors representing oppositional defiant, conduct problem, and inattentive behavior provided the best fit for the total sample (RMSEA ≤ 0.005, CFI ≥ 0.94, and TLI ≥ 0.92). Comparative fit index (CFI) is less sensitive to sample size compared to the chi-squared statistic, and values greater than 0.90 are considered to indicate reasonably good fit (Axberg et al., 2008). All items were found to load significantly on the general factor from $\lambda = 0.36$ to 0.66, and 10 out of the 22 items loaded more strongly on the specific factor than on the general factor. Specifically, the inattention factor demonstrated the strongest loadings, with four out of

four items loading more strongly on the specific factor compared to the general factor. Findings related to model fit were similar in the BPT and PMTO samples. Overall, the general factor was found to explain about half the variance in scores, indicating that the common variance was equally spread across general and specific factors.

Hukkelberg (2017) formulated several notable conclusions from the findings. First, the structure of the 22-item ECBI was best represented by a general factor of problem behavior and three uncorrelated specific factors of oppositional defiant, inattentive, and conduct problem behavior. Second, the results were comparable between the BPT and PMTO samples, suggesting that the structure is not dependent on the level of problem behavior and holds across varying sub-clinical groups. Lastly, Hukkelberg further evaluated the bifactor model using sources of variance in addition to model indices to address the notion that fit indices often favor bifactor models (Murray & Johnson, 2013). Findings demonstrated that the common variance was equally spread across general and specific factors. The results provided support for the bifactor model of the 22-item ECBI and consideration of the specific factors when using the ECBI in clinical and research settings.

The findings reported by Hukkelberg (2017) should be interpreted with caution due to several limitations, and additional research is necessary to replicate the findings before recommendations can be made to modify the interpretation of the ECBI. First, Hukkelberg (2017) used McDonald's ω as a measure of scale reliability. Although there is no consensus on the cut-off values for ω, Hukkelberg opted to use 0.30 as a cut-off for scale reliability and concluded that the majority of the subscales demonstrated adequate reliability (Reise, Bonifay, & Haviland, 2013). However, if a value of at least 0.50 were

used, only the inattentive scale in both the BPT and PMTO groups would have demonstrated adequate scale reliability (Hukkelberg, 2017). Further, ω for the general factor was high, suggesting that most of the variance was explained by the general factor. Therefore, an alternative interpretation of the values indicates that the majority of the variance was explained by the general factor and that an additive value across the items would provide adequate insight into conduct behavior problems as a whole.

In order to reconcile the two conflicting interpretations of Hukkelberg's (2017) findings, additional research is necessary to evaluate the use of a bifactor model for the ECBI data, and a consensus on the cut-off values for ω is necessary to assess the reliability of the scores for the specific factors. Based on the current findings alone, radical changes to the interpretation of the ECBI and application of a bifactor model are not supported. However, in certain instances, such as group assignment and categorizing by problem type, evaluating active components of treatment interventions and deciding between varying parent-training interventions, it may be helpful to specify a bifactor model to understand parent endorsement and its relationship to external factors more fully.

Ethnically diverse samples. The ECBI is available in Chinese, English, German, Japanese, Korean, Lebanese, Norwegian, Russian, Spanish, Dutch, and Swedish (Abrahamse et al., 2015; Eyberg & Pincus, 1999). Due to a paucity of psychometric findings and available norms for diverse cultures, independent investigators have evaluated the functioning of translated versions of the ECBI (Axberg et al., 2008; Ismaili, 2014; Leung, Sanders, Leung, Mac, & Lau, 2003; Reedtz, Bertelsen, Lurie, Hendegard, Clifford, & Morch, 2008; Rhee & Rhee, 2015). Overall, the ECBI demonstrated good

internal consistency as well as concurrent and discriminative validity across cultures. Conclusions regarding the recommended cutoff scores for the Problem and Intensity scales support identical cutoff scores for Swedish and Chinese populations, lower cutoff scores for Korean and Dutch samples, and higher cutoff scores for Spanish samples (Axberg et al., 2008; Leung et al., 2003; Reedtz et al., 2008; Rhee & Rhee, 2015). Aside from the studies using the Swedish and Korean versions of the ECBI, the majority of the investigations did not include an evaluation of the factor structure of the translated measure.

Axberg et al. (2008) and Rhee and Rhee (2015) included evaluation of the factor structure of the ECBI in their analytic procedures. Axberg et al. evaluated the 22-item ECBI and found the Burns and Patterson (2000) three-factor structure to fit the data adequately. Rhee and Rhee evaluated the complete 36-item ECBI and found eight meaningful factors using EFA. Notably, the seven items that loaded onto the "ADHD behavior" factor are identical to those identified by Burns and Patterson (2000). However, Rhee and Rhee did not pursue a CFA to confirm their findings.

The findings reported by Axberg et al. (2008) and Rhee and Rhee (2015) provide preliminary support for a multidimensional model of the ECBI in culturally diverse samples. In addition, the Burns and Patterson (1991, 2000) tripartite model has been replicated by at least two studies completed in the United States (Stern, 2007; Weis et al., 2005). These findings question the unidimensionality of the ECBI as supported by the original authors and subsequent investigations (Abrahamse et al., 2015, Butler, 2011; Colvin et al., 1999; Gross et al., 2007). Moreover, the unique findings related to normative data add to the notion that cultural factors impact response patterns, and the

underlying construct of measurement tools may be unique to different populations. Overall, the variability in findings, dearth of culturally diverse samples, and reliance on sample-specific analytic procedures provides a rationale for further investigation of the measurement equivalence of the ECBI in culturally diverse populations.

Present Study

Over the past decade, the cultural diversity of the United States population has significantly increased, and Hispanic individuals are the fastest growing minority group (U.S. Census Bureau, 2016). As the diversity of the general population continues to grow, the disparities experienced by ethnic minorities increase in significance. For example, problematic behaviors in children are a common reason for referral for mental health assessment and treatment (Visser et al., 2014). However, minority youth experience mental health disparities related to diagnosis, intervention, and access to care. Specifically, Hispanic children are diagnosed with ADHD at lower rates compared to their non-Hispanic European American counterparts; are less likely to receive quality mental health treatment; and have higher treatment drop-out rates (Morgan, Staff, Hillemeier, Farkas, & Maczuga, 2013; Olfson, Moitabai, Sampson, Hwang, & Kessler, 2009).

Several factors contributing to the disparities have been identified; however, a salient factor in mental health assessment is the limited number of valid and culturally sensitive assessment measures (Pumariega et al., 2005). Further, van de Vijver and Poortinga (2005) proposed that before a measure can be used in cross-cultural groups, structural and measurement equivalence must be established. As rating scales are the

most commonly used tools in assessment, investigation into the cross-cultural use of popular rating scales is a worthy area of focus.

The ECBI is a widely used screening measure for problematic behaviors in children and adolescents (Berkovits, O'Brien, Carter, & Eyberg, 2010; Funderburk et al., 2003). It is marketed and routinely used as a one-dimensional measure of disruptive behaviors in children and adolescents (Eyberg & Pincus, 1999). However, the ECBI was originally developed using data from predominately non-Hispanic European American samples, and subsequent research has largely lacked cultural diversity in the samples (Burns & Patterson, 1991, 2000; Colvin et al., 1999). Noteworthy are the discrepant results related to the factor structure of the ECBI. Multiple investigators have explored the factor structure of the ECBI in rather homogenous samples with limited inclusion of ethnic minority participants. The one-factor and three-factor structures have the most evidence in predominately non-Hispanic European American as well as international samples. However, the structural and measurement equivalence of the ECBI is essentially unknown in Hispanic populations and other ethnic minority groups. More research is necessary to make definitive conclusions about the factor structure of the ECBI and to provide recommendations related to its cross-cultural use.

In summary, additional exploration into the factor structure of the ECBI is warranted for several reasons. First, an alternative interpretation of ECBI data, such as the tripartite model, could increase the clinical and research utility of the ECBI. Similar to the analysis by Weis et al. (2005) and the model proposed by Burns and Patterson (2000), further evaluation of the three-factor model may provide additional support for the discriminative validity of the factors. For example, individual factor scores may be

useful throughout treatment by allowing the therapist to focus on specific problem areas and tailoring intervention. In research, group assignment may be more specific based on the proposed ECBI factors rather than total scores. Moreover, in program evaluation, factor-based interpretation of the ECBI can lead to more precise understanding of treatment effects, intervention efficacy, and active treatment components.

Second, a salient problem in the existing research is the limited ethnic diversity in the samples and the sample-dependent analytical methods used. The existing research focuses on mostly non-Hispanic European American samples with some inclusion of African Americans and miniscule incorporation of Hispanic individuals. As of 2015, the population of the United States was primarily European American (i.e., 76%), 13% African American, and 17% Hispanic (U.S. Census Bureau, 2016). However, Hispanics constitute the fastest growing minority population across the United States, and, in some parts of the U.S., such as South Florida, the Hispanic population ranges from 28% to 67%. Hispanic cultural values (e.g., *familismo, respeto,* child rearing practices) may impact the measurement and structural equivalence of rating scales. For example, collectivistic cultural values may facilitate a more accepting and understanding view of child behavior or may make Hispanic parents less likely to rate externalizing behaviors as problematic (Canino & Guarnaccia, 1997). Therefore, the current study includes a culturally diverse sample, with Hispanic participants representative of their proportion in the geographical region where the ECBI is being used.

Third, the available research relating to the factor structure of the ECBI utilizes primary CTT techniques and one identified use of IRT to confirm the single facture structure found by Eyberg and Ross (1983). Abrahamse et al. (2015) found support for a

one-dimensional factor structure using IRT methods. However, item level statistics, including discussion of DIF, were not reported. Additional analysis using IRT methods, specifically Rasch Modeling, is warranted to supplement the research and to explore further the psychometric properties of the ECBI.

The aim of the present study is to confirm the factor structure and to assess the psychometric functioning of the ECBI in an ethnically diverse sample. Both Confirmatory Factor Analysis and Rasch modeling were used.

Hypotheses. In the current study, the following hypotheses were tested:

1. The three-factor model of ECBI intensity scale will provide a better fit to the data than a one-factor model as demonstrated in other research (e.g., Axberg et al., 2008; Burns and Patterson, 1991, 2000; Weis et al., 2005).
2. Several ECBI intensity scale items (e.g., externalizing behaviors) will function differently between non-Hispanic and Hispanic samples.
3. The ECBI will demonstrate adequate reliability within an ethnically diverse sample.

Chapter 3: Methods

Prior to any data analysis, approval was obtained from the Institutional Review Board (IRB) at Nova Southeastern University to conduct archival research. All analyses were conducted using de-identified data.

Participants

Data from a de-identified archival database of an ADHD assessment and treatment clinic located in Broward County in South Florida were used for the study. The clinic is located in a university-based psychology services center and provides psychological services to an ethnically diverse population. According to the 2016 U.S. Census data, Broward County is 38% European American, 28% Hispanic, and 27% African American (U.S. Census Bureau, 2016). The data collected from the clinic database reflected the ethnic diversity of the area allowing for novel findings relating to the dimensionality and overall psychometric functioning of the ECBI in an ethnically diverse sample and, in particular, a Hispanic sub-population. The secure database contains de-identified client information, including gender, age, diagnosis, and assessment scores.

Children are referred to the clinic for assessment and treatment of a wide range of childhood problems, including mood, behavior, and learning difficulties. As part of the assessment and treatment process, caregivers are asked to complete several parent-report measures of behavior and mood functioning, including the ECBI. Children whose parent(s) have completed the ECBI and are between two and 16 years old were included in the data subset for analysis. Individual item scores and the Intensity scale scores from the ECBI were used in the analysis. Scores from additional measures of behavior, such as

the Conners Parent Rating Scale, Third Edition (Conners, 2008) was used to assess convergent and discriminant validity of the ECBI scale(s) in the present study. In addition, demographic information, including ethnicity, gender, and age of the participants was used in the analysis.

Sample Characteristics

The total sample included 221 children and adolescents whose mothers completed the ECBI. The sample was 72% male and 28% female, with an average age of 9.32 years (range 3-17, *SD* = 2.936). A total of 147 (66.8%) of the parents who responded about their children were married; 37 (16.8%) were divorced; seven (3.2%) were separated; 22 (10.0%) were single and had never been married; five (2.3%) were living with someone; two (<1%) were widowed; and one respondent's marital status was missing. Related to ethnicity, 43.4% of the sample was Hispanic, 41.2% was European American, 12.2% was African American, and 3.2% identified as "other." Of the total sample, 194 parents provided yearly household income information, and 27 did not. A total of 64 (33.0%) reported a yearly household income of over $70,000; 18 (9.3%) reported between $60,000 and $69,999; 13 (6.7%) reported between $50,000 and $59,000; eight (4.1%) reported between $40,000 and $49,000; 30 (15.5%) reported between $30,000 and $39,999; 44 (22.7%) reported between $20,000 and $29,999; 15 (7.7%) reported between $10,000 and $19,999; and two (1.0%) reported a yearly household income of less than $10,000.

Measures

The Eyberg Child Behavior Inventory. The ECBI is a 36-item parent-report measure designed to assess conduct problems in both children and adolescents (Eyberg &

Ross, 1978). The ECBI has two scales, a Problem scale and an Intensity scale. For the Intensity Scale, caregivers are asked to indicate the severity of each of the 36 behaviors by rating the frequency of occurrence on a seven-point Likert-type scale. The scale ranges from a value of one, indicating "*never,*" to seven, indicating "*always.*" For the Problem scale, caregivers are asked to indicate in a yes or no format whether they consider the particular behavior to be problematic regardless of the intensity. Caregiver ratings above a score of 131 on the Intensity Scale and 15 on the Problem Scale are considered to indicate problems within the clinically significant range. The ECBI has demonstrated good internal consistency, test-retest reliability, and sensitivity to treatment effects (Colvin et al., 1999; Eyberg & Robinson, 1983; Robinson et al., 1980).

Conners Parent Rating Scale, Third Edition. The initial Conners Parent Rating Scale (CPRS; Conners, 1970) was developed as a comprehensive checklist designed to gather caregiver report of problematic behaviors in children. The revision of the CPRS (Conners, Sitarenios, Parker, & Epstein, 1997) resulted in a similar factor structure, including seven dimensions: Cognitive Problems, Oppositional, Hyperactivity-Impulsivity, Anxious-Shy, Perfectionism, Social Problems, and Psychosomatic subscales. The CPRS-Revised (CPRS-R) included fewer items (i.e., 57) while providing a more comprehensive assessment of ADHD-related behaviors. Coefficients for six-week test-retest reliability range from 0.42 to 0.78 for the majority of scales, although only 0.13 for the Social Problems Scale, and acceptable internal consistency was obtained, with values ranging from 0.75 to 0.94.

The CPRS was again revised, resulting in the Conners Parent Rating Scale, 3rd edition (Conners-3P; Conners, 2008). The Conners-3P consists of 110 items and is a

narrow-band caregiver report measure of ADHD and other related disorders as well as oppositional/defiant and problematic conduct behaviors. The items and scales of the Conners-3P are similar to established diagnostic criteria and, much like its predecessor, scoring yields index values indicative of problem areas and possible clinical syndromes. Test-retest reliability coefficients for two- to four-week administrations range from 0.70 to 0.98, and internal consistency values range from 0.77 to 0.98.

Analytic Procedure

Descriptive statistics were reported for the sample, such as gender, age, and ethnicity. Then, the following steps were used to evaluate the psychometric properties of the ECBI between Hispanic and non-Hispanic groups. First, the dimensionality of the ECBI was explored using confirmatory factor analysis methodology. Second, a model appropriate for polytomous data, the Rating Scale Model (RSM; Andrich, 1978), was employed to evaluate the rating scale functioning of the scales that comprise the ECBI. Third, several key aspects of the RSM were evaluated further, including dimensionality, item fit, person fit, and reliability. Fourth, the degree to which the items function similarly across Hispanic and non-Hispanic groups was evaluated. Fifth, convergent and discriminant validity were assessed. The analytic process was iterative in nature, such that the steps initially planned were modified depending on the results of each phase. All descriptive statistics were calculated using IBM SPSS version 25 (IBM Corp., 2017), while the psychometric analyses were performed using IBM SPSS AMOS version 25 (Arbuckle, 2017) and WINSTEPS version 4.0.1 (Linacre, 2017).

Dimensionality. There are conflicting results regarding dimensionality of the ECBI in the existing literature. Specifically, some authors have argued for a one-factor

model (Abrahamse et al. 2015; Colvin, Eyberg, & Adams, 1999) while other authors have suggested a three-factor model (Burns &Patterson, 2000). Therefore, based on existing literature, two nested CFA models (one-factor versus three-factor) were fit to the data to explore the dimensionality of the ECBI items. Fit criteria included absolute fit indices, e.g., chi-square and standardized root mean square residual (SRMR), parsimony corrected indices, e.g., root mean square error of approximation (RMSEA), and comparative fit indices, e.g., comparative fit index (CFI). In order to estimate fit, criteria suggested by Hu and Bentler (1999) were used as general guidelines (i.e., CFI $\geq$ 0.95, SRMR $\leq$ 0.08. and RMSEA $\leq$ 0.06).

Rating scale functioning. An RSM was employed to examine the rating scale functioning of the scales that comprise the ECBI. The RSM is a member of the family of Rasch models that is intended for use with polytomous items (Andrich, 1978). It has the same features as the Rasch model, such as unidimensionality (i.e., a set of items measure one latent trait) and local independence of items (i.e., test items are independent of one another), in addition to similar specifications such as equal item discrimination and monotonicity (Embertson & Reise, 2000). Further, RSM is advantageous in that a person's response to an item is governed by his or her report of the latent trait and the RSM's one parameter, i.e., the item's difficulty. For dichotomous items, item difficulty is the trait level required for a respondent to have a 50-50 probability of endorsing the item (Furr, 2018). An item with a high difficulty level requires a higher trait level to endorse, while a less difficult item will require a lower trait level. Item difficulties have a mean of zero and a standard deviation of one. For polytomous items, such as in RSMs, each item has C-1 thresholds, where C = number of response option categories. Item difficulty is

the mean of the thresholds. It is important to note that since the items were found to share a similar rating scale structure, a RSM, rather than the originally planned Partial Credit Model (PCM; Masters, 1982) was used.

Item discrimination, or slope, is the degree to which an item differentiates individuals who have high trait levels from those with low trait levels (Embretson & Reise, 2000; Furr, 2018). An item with a high discrimination value indicates a strong relation to the underlying trait measured, while low discrimination (e.g., value of 0) indicates that the item is unrelated to the underlying trait. In RSM, all items are assumed to discriminate equally, and item difficulty is the characteristic, or parameter, estimated in order to understand the probability that the person will respond in a particular way to a response option.

Another specification of the RSM is monotonicity, which indicates that as trait level increases, e.g., severity of problematic behavior, so does the probability of endorsing an item, e.g., selecting a higher item severity such as "strongly agree" (Embretson & Reise, 2000; Linacre, 2017). If the assumption of monotonicity is violated, the rating scale may be disordered. To test this assumption, Andrich thresholds, i.e., ordered versus disordered (Andrich, 2006), were used. An Andrich threshold, also referred to as step difficulty, is defined as the trait level at which one has an equal probability of endorsing adjacent response options, e.g., the respondent has an equal probability of endorsing "strongly disagree" and "disagree" (Andrich, 2006, Embretson & Reise, 2000; Linacre, 2017). RSM does not allow for the Andrich thresholds of the ratings scales to vary between items and it was expected that for each item the threshold would increase with category value along the rating scale. Disordered thresholds

suggested that the rating scale, e.g., “strongly disagree” to “strongly agree”, was not being consistently interpreted in an ordered fashion across respondents. As the Andrich thresholds were found to be disordered and monotonicity was violated, combining specific response options was utilized to help improve rating scale functioning.

Additionally, a PCA of the probability residuals derived from the RSM was conducted in order to further assess dimensionality of each scale (Bond & Fox, 2015; Linacre, 2017). The variance explained by the “Rasch dimension,” and the variance attributed to the standardized residuals after the Rasch dimension had been accounted for, was used to assess dimensionality. At least 40% of the variance explained by measures and less than 15% of the total unexplained variance accounted for by the first contrast were tentative guidelines for interpreting dimensionality. Additionally, an eigenvalue less than two for unexplained variance in the first contrast suggested that there was likely only one meaningful dimension explaining responses (Linacre, 2017; Raîche, 2005). Further, a factor sensitivity ratio (Allison, Baron-Cohen, Stone, & Muncer, 2015; Bond & Fox, 2015; Wright & Stone, 2004) was used to determine to what extent the measure was influenced by the unexplained variance. While there is not a suggested cutoff or threshold for interpretation for this ratio, it can be, and was, used to evaluate the percentage of the measure that is impacted by unexplained relationships between items. This index was generated by dividing the residual variance (variance unexplained by the Rasch dimension) eigenvalue units by the Rasch dimension variance (variance explained by the Rasch dimension) eigenvalue units. This ratio can be multiplied by 100 to generate a percentage of the measure that is affected by the unexplained relationships between items.

Model fit. Dependent on the dimensionality results, it was appropriate to fit the items of the three proposed scales individually to the RSM rather than fitting a one factor model. To estimate the fit of the data to the model, item polarity, the observed average measures for persons, and mean square statistics (i.e., outfit and infit) for both persons and items were considered when determining item response functioning and model fit. Item polarity shows the degree to which items align with the latent variable (Bond & Fox, 2015). Polarity is estimated using point-measure correlations as reported by WINSTEPS and is related to the fundamental assumption that higher ability aligns with higher ratings on items and vice versa (Linacre, 2017). Positive point-measure correlations suggested that higher response options aligned with higher levels of the intended construct.

The category functioning for each item was further assessed by visual inspection of the observed average measures plot for each scored category in order to confirm that higher trait levels resulted in endorsing a higher category (Linacre, 2017). It was expected that the category observed averages for each item would ascend from the left to the right of the plot. Furthermore, the item hierarchy, the spread of the person sample, and the item range was examined.

Person fit (i.e., the extent to which individual responses differ from the model expectations) and item fit (i.e., the extent to which each item functions differently than expected within the model) statistics were also evaluated. A mean square fit statistic value close to one and within the range of 0.5 to 2.0 suggested adequate model fit (Linacre, 2017). In order to investigate item bias that may exist between groups, the presence of DIF between non-Hispanic and Hispanic respondents was used in order to assess the cross-cultural functioning of the items of the ECBI.

DIF is useful in understanding differences in test scores in cross-cultural assessment (Cauffman & MacIntosh, 2006; Tennant et al., 2004). For a unidimensional scale, group membership is not expected to influence response patterns significantly. The presence of DIF suggests that group membership, in addition to the respondent's standing on the latent trait and the item's difficulty, may be influencing the probability of endorsing an item (Furr, 2018). If the item's functioning is significantly impacting responses, items can be re-worded; removed; or, if the differential item functioning is cancelled out by another item, left unaltered. Significance testing was used to assess DIF (i.e., pair-wise comparison of the item difficulties in two groups, non-Hispanic vs. Hispanic) and items with a p-value below .05 ($p < .05$) were considered to show statistically significant DIF (Linacre, 2017). Additionally, examination of the DIF contrast, i.e., the difference in DIF between each of the comparisons (i.e., Hispanics compared individually with each other ethnic group), was used to assess DIF (Linacre, 2017). |DIF contrast| ≥ 0.5 logits (Linacre, 2017) was considered to be substantive.

Reliability. Reliability of the ECBI was assessed using Rasch-based estimates of reliability including separation coefficients and reliability indices both for persons and for items (Bond & Fox, 201; Linacre, 2017). The separation coefficient is a "signal-to-noise" ratio where signal is the true variance and noise is the error variance (Linacre, 2017). Separation values less than two for persons suggest that the instrument may not be sensitive enough to differentiate between high and low responders, and values less than three for items suggest that the sample may not be large enough to confirm the order of the item difficulties. Reliability is the extent of reproducibility of the order of person and of item measures, and is the true variance divided by the observed variance (where

observed variance = true variance + error variance). Reliability indices have a range of zero to one, and values less than 0.5 imply high measurement error. Notably, the Person Reliability index is analogous to Cronbach's alpha, while the Item Reliability index has no classical test theory equivalent. Cronbach's alpha was used to further assess reliability. Reliability indices and Cronbach's alpha values equal to or greater than 0.7 were deemed acceptable (Linacre, 2017).

Validity. Finally, evidence of the validity of the ECBI scores was assessed. The Standards for Educational and Psychological Testing provides an updated view of validity that emphasizes construct validity as a principal concept of validity (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014; Downing, 2003). The five facets related to the construct validity of a test include the test content, the internal structure of the test, the interpretation and processes involved in responding (i.e., response processes), the association of test scores with other variables (i.e., convergent and discriminant validity), and the consequences of test use (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014; Downing, 2003; Furr, 2018). Of these five facets of validity, the internal structure as well as convergent and discriminant validity of the ECBI was explored. Consequences of test use is an important aspect of validity that relates to the effects, both unintended and intended, of using a measure. However, consequential validity is beyond the scope of this study. Further, although test content and response processes were not evaluated, research related to these facets of validity for the ECBI is reviewed below.

The internal structure of the ECBI is related to the dimensionality, which was evaluated as described in the analytic section above. With respect to associations of test scores with other variables (i.e., convergent, discriminant validity), the ECBI has been found to be highly correlated with other parent-report measures of behavior (Gross et al., 2007). The association with other variables was assessed by testing the extent to which the ECBI correlates with other measures of related constructs (i.e., convergent validity), such as specific subscales of the Conners Parent Rating Scale (Conners, Sitarenios, Parker, & Epstein, 1997; Furr, 2018). The ECBI scores were expected to be highly correlated with various subscales of the CPRS, such as the Inattentive and Oppositional behavior subscales. In order to assess the discriminant validity of the ECBI test scores, the extent to which ECBI scores are correlated with theoretically unrelated variables, such as the Peer Relations subscale of the CPRS, was explored. ECBI test scores were expected not to be highly correlated with the Peer Relations subscale.

Related to the test content of the ECBI, in a focus group study a cohort of African American, Latino, and non-Latino parents found the items of the ECBI to be relevant indicators of child behavior problems and to represent an adequate range of content related to child behavior problems (Sivan, Ridge, Gross, Richardson, & Cowell, 2008). Specifically, 70 parents were asked to list behaviors that a child that they viewed as "problematic" would exhibit. Comparison of parent responses to the ECBI items indicated that the ECBI items were good markers of parent-reported problematic behaviors. Notably, when asked what behaviors were not included in the ECBI and should have been, parents identified internalizing behaviors (e.g., "passive or withdrawn" and "looks unhappy/won't smile") and externalizing behaviors (e.g., "bites" and "cruel or

abusive to animals"). However, none of the missing indicators of child behavior problems were reported in more than three of the 15 focus groups, and there were no notable patterns in the parent's racial/ethnic background and the behaviors identified as missing.

Related to the interpretation of items, in the same focus group study by Sivan and colleagues (2008), parents were asked to identify any words or items of the ECBI that they did not understand or that people they knew might not understand and to highlight any words or phrases that may be upsetting or culturally biased. None of the ECBI items were identified as culturally biased or upsetting, and the majority of the items were deemed understandable by the participants. However, some items were identified as either too vague or as a behavior that could be construed as normal (e.g., "dawdles or lingers at mealtime," "refuses to do chores when asked," and "interrupts"). Nevertheless, the authors concluded that, overall, the items of the ECBI were understandable and similarly interpreted by the parents in the focus groups.

Beyond the study by Sivan et al. (2008), the item selection method utilized by the authors of the ECBI, as well as the rationale for the wording of the items, provides further support for the test content and response process validity of the ECBI (Eyberg & Ross, 1978; Sivan et al., 2008). The ECBI authors selected the behaviors listed in the ECBI from relevant clinical cases in order to represent the most commonly parent-reported behavior problems (Eyberg & Ross, 1978). In order to limit the potential for discrepancies in the interpretation of the behaviors, the authors chose specific behavioral descriptors, e.g., "refuses to do chores when asked," rather than general descriptive terms, e.g., "is defiant." Therefore, although the test content and response process validity of the

ECBI was not explored in this study, the ECBI has support for both of these facets of construct validity.

Chapter 4: Results

Descriptive Statistics

Descriptive statistics are displayed in Table 5. The data were normally distributed as indicated by skewness and kurtosis values close to zero and within the range of -2 to 2 (Byrne, 2010; George & Mallery, 2010).

Table 5

Descriptive Statistics of the ECBI Total Score

Group	N	ECBI Mini.	ECBI Max.	ECBI Mean	Std. Deviation	Skewness		Kurtosis	
						Statistic	SE	Statistic	SE
Total	221	45	213	123.26	36.5	-0.085	0.164	-0.719	0.326

Note: ECBI= Eyberg Child Behavior Inventory; Mini = Minimum raw score; Max. = Maximum raw score; Std. Deviation = Standard Deviation; SE = Standard Error

Hypothesis One

The first hypothesis states that a three-factor model would provide a better fit to the data than a one-factor model as demonstrated in prior research (e.g., Axberg et al., 2008; Burns & Patterson, 1991, 2000; Weis et al., 2005). In order to test this hypothesis, CFA of a one- and a three-factor model was employed to explore the dimensionality of the ECBI. Based on the research by Burns and Patterson (2000), the factor analyses were performed using 22 intensity scale (IS) items out of the ECBI's original 36 items. The 14 ECBI items not included in the analysis were identified by Burns and Patterson to have low factor loadings, to be conceptually different from the majority of the other items loading on that factor, and/or to be indicative of a meaningless dimension. The three meaningful factors that emerged from the Burns and Patterson EFA were Oppositional Defiant Behavior Toward Adults (ODBTA), Inattentive Behavior (IB), and Conduct Problem Behavior (CPB). For the one factor CFA model, the items of each of the three

meaningful factors were combined into a single, 22-item factor representing general problematic behaviors, as demonstrated by Burns and Patterson. A list of the factors and their items, as well as the factor loadings for the four-factor model from the EFA by Burns and Patterson, can be found in Table 4.

Determination of model fit for the one- and three-factor CFAs was established based on comparison of multiple fit indices, i.e., χ^2, CFI, SRMR, and RMSEA, of the one- and the three-factor models using pre-established fit criteria, i.e., CFI ≥ 0.95, SRMR ≤ 0.08, and RMSEA ≤ 0.06 (Hu & Bentler, 1999). Maximum likelihood estimation was used for the CFAs.

The one-factor model resulted in a poor fit, χ^2 (209) = 1074.547, $p < .001$. The CFI, SRMR, and RMSEA values were 0.666, 0.115, and 0.137, respectively. The three-factor model also resulted in a poor fit, χ^2 (206) = 567.946, $p < .001$. The CFI, SRMR, and RMSEA values were 0.860, 0.071, and 0.089, respectively. However, the modification indices, which indicate the change in the overall chi-square if the parameters of the model were changed, suggested that correlating three pairs of error terms would improve model fit. The three item pairs were "verbally fights with sisters and brothers" correlated with "physically fights with sisters and brothers" and "steals" with "lies" from the CPB factor, and "has temper tantrums" with "yells or screams" from the ODBTA factor. Given the similar content of the item pairs (i.e., discord between siblings; deceitful behaviors; and disruptive behaviors) and the likelihood of the item pairs sharing a unique secondary dimension (e.g., whether the child has siblings or not for items 27 and 25), correlating their error terms is justifiable (see Brown, 2015). It is also theoretically reasonable, as the suggested correlations do not cross factors and Burns and

Patterson (2000) also correlated errors for the CPB item pairs. Therefore, both the one- and three-factor models were re-specified with three correlated errors.

The one-factor model with three correlated errors resulted in an improved, but still a poor fit, χ^2 (206) = 864.999, $p < .001$. The CFI, SRMR, and RMSEA values were 0.746, 0.106, and 0.121, respectively. The three-factor model with three correlated errors resulted in an acceptable fit, χ^2 (203) = 385.032, $p < .001$. The CFI, SRMR, and RMSEA values were 0.930, 0.060, and 0.064, respectively. In addition to the fit indices reported, the chi-square difference test indicated that the three-factor model with three correlated errors provided a significantly better fit than the one-factor model with three correlated errors, $\Delta\chi^2(3) = 479.967$, $p < .001$. Additionally, the Akaike Information Criterion (AIC) index for the three-factor model with three correlated errors resulted in the smallest value across all the models tested, AIC = 485.032, indicating that this model is the best fit for the data. The detailed results of the one- and three-factor CFAs, as well as the results of the model comparisons using the likelihood ratio chi-square test and AIC index can be found in Table 6.

Table 6

Model Fit Indices of the One- and Three-Factor Models with Varying Correlated Error Terms for the Seven-point Rating Scale Structure

Model	χ^2	df	χ^2 /df	CFI	SRMR	RMSEA	AIC
No Correlated Errors							
1-Factor	1074.547*	209	5.141	0.666	0.1153	0.137	1162.547
3-Factor	567.946*	206	2.757	0.860	0.0714	0.089	661.946
One Correlated Error							
1-Factor	928.259*	208	4.463	0.722	0.1085	0.125	1018.259
3-Factor	444.257*	205	2.167	0.908	0.0614	0.073	540.257
Two Correlated Errors							
1-Factor	892.861*	207	4.313	0.736	0.1068	0.123	984.861
3-Factor	413.407*	204	2.027	0.919	0.0614	0.068	511.407
Three Correlated Errors							
1-Factor	864.999*	206	4.199	0.746	0.1063	0.121	958.999
3-Factor	385.032*	203	1.897	0.930	0.0609	0.064	485.032

*$p < .001$

The standardized regression coefficients for the ODBTA, IB, and CPB factors for the three-factor model with three correlated error terms are presented in Table 7. The correlation coefficients between the factors and Cronbach's alpha for each factor are also included.

Table 7

Standardized Regression Coefficients of the Three-Factor Model

Items	Factors		
	ODBTA	IB	CPB
Oppositional Defiant Behavior Toward Adults			
11. Argues with parents about rules	0.833		
10. Acts defiant when told to do something	0.850		
9. Refuses to obey until threatened with punishment	0.837		
14. Sasses adults	0.753		
5. Refuses to do chores when asked	0.621		
12. Gets angry when doesn't get own way	0.854		
8. Does not obey house rules on own	0.757		
7. Refuses to go to bed on time	0.495		
13. Has temper tantrums	0.748		
17. Yells or screams	0.710		
Inattentive Behavior			
31. Has short attention span		0.872	
30. Is easily Distracted		0.831	
34. Has difficulty concentrating on one thing		0.750	
32. Fails to finish tasks or projects		0.726	
Conduct Problem Behavior			
25. Verbally fights with sisters and brothers			0.360
27. Physically fights with sisters and brothers			0.443
23. Teases or provokes other children			0.645
24. Verbally fights with friends his or her own age			0.738
26. Physically fights with friends his or her own age			0.618
22. Lies			0.498
21. Steals			0.332
19. Destroys toys and other objects			0.584

Note: The correlation between the ODBTA factor and the IB factor was 0.34. The correlation between the IB factor and the CPB factor was 0.32. The correlation between the CPB factor and the ODBTA factor was 0.70. Cronbach's alpha was 0.927 for the ODBTA scale, 0.873 for the IB scale, and 0.778 for the CPB scale.

Hypothesis Two

The second hypothesis states that several of the ECBI IS items (e.g., externalizing behaviors) will function differently between non-Hispanic and Hispanic groups. In order to test this hypothesis, a Rasch-based model appropriate for use with polytomous data, such as the Masters Partial Credit Model (Masters, 1982; PCM) or the Andrich Rating Scale Model (Andrich, 1978; RSM), was used to explore the psychometric properties of the ECBI scales.

Since the three-factor model emerged as the superior fitting model using CFA, the psychometric functioning of each of the three scales was explored individually. The psychometric evaluation of each scale involved several steps, including evaluation of the rating scale functioning; the dimensionality of the scales; the item and the person fit; the differential item functioning, i.e., the degree to which the items function similarly across Hispanic and non-Hispanic groups; the reliability of the scales within an ethnically diverse sample (hypothesis three); and, finally, the validity of the ECBI scale scores.

Two models appropriate for polytomous data, the PCM (Masters, 1982) and the RSM (Andrich, 1978), were considered for this analysis. The PCM allows for each item to have a unique rating scale structure (Masters, 1982) and is ideal for scales whose items do not share the same rating scale structure (Linacre, 2017). Alternatively, an RSM is one in which all the items of the scale share the same rating scale structure (Andrich, 1978; Linacre, 2017). Both models are considered to be within the Rasch family of measurement models in that the person's ability, or trait level, and the item's difficulty are the parameters that predict the probability of endorsing a response category. For this study, the terms "person ability" and "trait level" refer to the severity rating of the child's behavior as reported by their mother. While the PCM was initially considered for this analysis, the RSM was ultimately chosen because all of the items across the three ECBI scales share the same rating scale structure and because it was ideal to keep the rating scale structure the same across all items for practicality of administration and scoring.

ODBTA: Rating scale functioning. The 10 IS items of the ODBTA scale, Table 8, were used for this analysis. Each item shared the same seven-point rating scale response structure. Response options ranged from "never" (i.e., one); "seldom" (i.e., two

or three); “sometimes” (i.e., four); “often” (i.e., five or six); to “always” (i.e., seven). As noted above, while a PCM was initially planned for this analysis, the RSM was ultimately selected given the shared rating scale structure of the items.

Table 8

Items of the ODBTA ECBI Scale

11.	Argues with parents about rules
10.	Acts defiant when told to do something
9.	Refuses to obey until threatened with punishment
14.	Sasses adults
5.	Refuses to do chores when asked
12.	Gets angry when doesn’t get own way
8.	Does not obey house rules on own
7.	Refuses to go to bed on time
13.	Has temper tantrums
17.	Yells or screams

Note: ODBTA= Oppositional Defiant Behavior Toward Adults

A specification of the RSM is the assumption of monotonicity. This reflects the expectation that as trait level, e.g., severity of problematic oppositional and defiant behavior, increases, so does the probability of endorsing a higher response category, e.g., selecting a higher item severity such as “always” (Embretson & Reise, 2000; Linacre, 2017). Violation of this specification would suggest that the categories may be disordered and are not functioning in accordance with the model.

Monotonicity for the ODBTA rating scale was assessed by examining the ordering, or disordering, of category observed averages and the ordering, or disordering, of Andrich thresholds (Andrich, 2006). The observed averages for a category reflect the average abilities of the respondents who endorsed that category (Linacre, 2017). In other words, observed averages reflect the average ability, or trait, levels needed for a respondent in this sample to endorse a certain response option (Linacre, 2017). Similar to Andrich thresholds, the RSM expects that observed averages will increase as response

options advance (Embreston & Reise, 2000; Linacre, 2017). Ordered observed averages are an indication that trait level advances as categories advance. Andrich thresholds, or step difficulties, are the point at which there is a 50-50 probability of endorsing adjacent response options and are indicated by the point at which response probability curves intersect for adjacent response options (Linacre, 2017). For polytomous items, there are *k*-1 thresholds, where *k* is the number of category options. It is expected that thresholds advance as trait level increases. Additionally, when thresholds are ordered, each category is most probable at some point along the scale.

In addition to monotonicity, category mean-square fit statistics, outfit and infit, were evaluated to assess category usage (Linacre, 1995). Mean-square statistic values near 1.0 indicate appropriate category usage (Linacre, 2017). Values greater than 2.0 indicate unpredictability in category usage that can distort measurement, and values less than 0.5 indicate overly predictable usage.

The results in Table 9 show the observed averages, Andrich Thresholds, and category mean-square fit statistics for the ODBTA rating scale items with seven response options. The results show ordered observed averages and mean-square values within the expected range, but disordered thresholds. This suggests that the scale is not functioning as expected and may indicate that some categories reflect narrow intervals of the latent variable (Linacre, 2017). Disordered thresholds often, but not always, violate the assumption that as trait level increases, the probability of endorsing the next response option increases smoothly along the scale. Specifically, for the ODBTA items, the threshold between the response options two and three is greater than the threshold

between response options three and four, and the threshold between the response options four and five is greater than the threshold between response options five and six.

Table 9

ODBTA Seven-Point Rating Scale Functioning

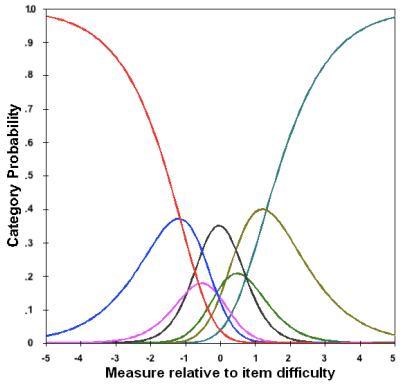
	Response Options						
	1	2	3	4	5	6	7
Andrich Threshold	None	-1.16	-0.10	-0.95	0.74	0.15	1.32
Observed Averages	-1.43	-0.85	-0.38	-0.06	0.26	0.63	1.19
Infit MNSQ	1.23	0.88	0.87	0.88	1.01	1.04	1.06
Outfit MNSQ	1.20	0.96	0.83	0.86	1.12	1.13	1.11

Note: MNSQ= Mean-square statistic.

The disordering of thresholds is visually depicted by the category probability curve for item 11 in Figure 1. These curves show the probability of endorsing each category. The Andrich thresholds are the points at which adjacent curves intersect. As each item shares the same rating scale structure for RSMs, each item has the same category probability curve structure.

Figure 1

Category Probability Curve for Item Eleven of the ODBTA Seven-point Scale



Note: Red = category one probability. Blue = category two probability. Pink = category three probability. Black = category four probability. Green = category five probability. Olive = category six probability. Teal = category seven probability.

When thresholds are found to be disordered, adjustments to the scale are needed to help the rating scale conform to model expectations (Embreston & Reise, 2000; Linacre, 2017). In such cases, collapsing adjacent categories can aid in improving rating scale functioning by addressing disordered thresholds (Embreston & Reise, 2000). Additionally, threshold disordering is often observed when categories correspond to a narrow interval of the latent variable (Linacre, 2017), which can be argued is the case for categories two and three, and five and six of the ODBTA rating scale. In fact, categories two and three are both labeled "seldom" and categories five and six are both labeled "often." Therefore, combining categories two and three, and categories five and six, and then re-assigning point-scores to the combined categories is reasonable, given the limited differentiation in their descriptions.

The responses for the ODTBA factor items were recoded using WINSTEPS from a seven-point rating scale to a five-point rating scale. The resulting response options for these items were "never" (i.e., one); "seldom" (i.e., two being combined categories two and three); "sometimes" (i.e., three); "often" (i.e., four being combined categories five and six); to "always" (i.e., five). The data were then re-evaluated using the RSM in order to assess the rating scale functioning. Specifically, observed averages, Andrich thresholds, and mean-square fit statistics were examined for the collapsed five-point rating scale. The results of the five-point rating scale are presented Table 10. The observed averages are ordered, and the mean-square fit statistics are within the expected range of 0.5 to 2.0. Additionally, the Andrich thresholds are ordered. These results show that for the ODBTA factor, the five-point rating scale is functioning in accordance with the RSM specifications.

It should be noted, that in order to confirm the findings regarding the superior fit of the three-factor model (hypothesis one) for the five-point rating scale structure, a CFA was performed with the rescored data for the 22-items of the ECBI. The results can be found in Appendix A.

Table 10

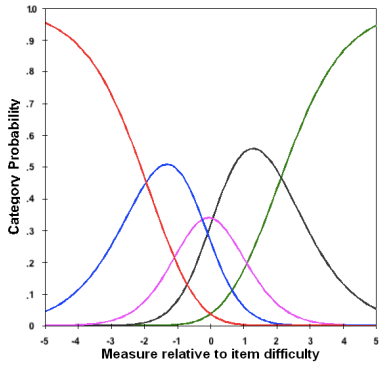
ODBTA Five-point Rating Scale Functioning

	Response Options				
	1	2	3	4	5
Observed Averages	-1.98	-0.95	-0.09	0.68	1.70
Andrich Threshold	None	-1.95	-0.25	-0.06	2.13
Infit MNSQ	1.20	0.87	0.85	1.00	1.04
Outfit MNSQ	1.19	0.89	0.82	1.09	1.08

Finally, visual examination of the category probability curves for the ODBTA items confirm that for the RSM with a five-point rating scale, as trait level increases, so will the probability of endorsing a more advanced category. The category probability curve for item 11 of the ODBTA scale is shown in Figure 2.

Figure 2

Category Probability Curve for Item Eleven of the ODBTA Five-point Scale



Note: Red = category one probability. Blue = category two probability. Pink = category three probability. Black = category four probability. Green = category five probability.

ODBTA: Dimensionality. In order to assess the dimensionality of the ODBTA scale, a Principal Components Analysis (PCA) of the probability residuals was used (Bond & Fox, 2015; Linacre, 2017). The RSM assumes unidimensionality, which indicates that all the items of the test, or, in this case, the ODBTA scale, measure one underlying construct, i.e., oppositional defiant behavior towards adults (Bond & Fox, 2015). It is expected that the Rasch dimension, i.e., the person measures and the item measures, will explain the majority of the variance in the data. Therefore, in order to assess dimensionality, the raw variance explained by measures, or the variance that can be explained by the Rasch measures, was evaluated. At least 40% of the variance explained by measures was tentatively used to evaluate unidimensionality (Linacre, 2017). The results of the PCA of probability of residuals for the ODBTA factor are presented in Table 11. The raw explained variance was 57.7%, and above the established guideline of 40%, suggesting that the ODBTA scale is unidimensional enough to meaningfully measure oppositional behaviors.

Table 11

Results of the PCA of Residuals of the ODBTA Scale

	Eigenvalue	Percentage
Total Raw Variance Explained by Measures	13.65	57.70%
Raw Variance Explained by Persons	7.76	32.80%
Raw Variance Explained by Items	5.88	24.90%
Total Unexplained Variance	10.00	42.30%
Unexplained Variance in 1st Contrast	2.02	8.60%
Unexplained Variance in 2nd Contrast	1.58	6.70%
Unexplained Variance in 3rd Contrast	1.25	5.30%
Unexplained Variance in 4th Contrast	1.09	4.60%
Unexplained Variance in 5th Contrast	0.95	4.10%

In order to assess further the dimensionality of the ODBTA scale, the raw variance unexplained was also used to determine whether another meaningful dimension, after the Rasch dimension had been accounted for, explained a significant amount of the residual variance (Bond & Fox, 2015; Linacre, 2017). The presence of a meaningful dimension would suggest multidimensionality in the data. The unexplained variance was explored using a PCA of the residual variance after the Rasch dimension has been accounted for. If the data were unidimensional, the components, or factors, identified by the PCA would be expected to be at "noise" level (Linacre, 1998, 2017; Wright, 1996). Therefore, less than 15% of unexplained variance in the first contrast along with an eigenvalue less than two were tentative guidelines for assessing unidimensionality. The unexplained variance accounted for by the first contrast was 8.6% with an eigenvalue of 2.02 (Table 11).

Since the eigenvalue for the first contrast was slightly above the expected value of two, examination of the content or the wording of the items at the top of the contrast table with the items towards the bottom of the contrast table aided in further clarifying the dimensionality of the ODBTA scale. The summary of residual loadings (Figure 3) shows items at the top of the table appear to be more related to externalizing behaviors, and items toward the bottom of the table appear to be more related to noncompliance and defiance. Given that the ODBTA scale is intended to measure oppositional defiant behaviors and that all the clusters of items reflect general oppositional defiant behaviors, it would not be conceptually sound to split the items into two dimensions.

Figure 3

Contrast Plot of the ODBTA Items' Residual Loadings

CON TRA	CL US	LOADING	MEASURE	INFIT MNSQ	OUTFIT MNSQ	ENTRY NUMBER	Item
1	1	.62	.34	1.01	1.06	A 7	13. Has temper tantrums.
1	1	.56	-.56	.74	.72	B 3	12. Gets angry when doesn't ge
1	1	.46	.20	1.07	1.05	C 8	17. Yells or screams
1	2	.32	.45	1.13	1.07	D 6	14. Sasses adults
1	2	.24	-.18	.83	.83	E 5	11. argues with parents about
1	3	-.55	-.11	1.12	1.24	a 9	5. Refuses to do chores when a
1	3	-.53	.01	.84	.94	b 4	8. Does not obey house rules o
1	3	-.53	-.04	1.78	1.84	c 10	7. Refuses to go to bed on tim
1	3	-.32	-.29	.83	.80	d 2	9. Refuses to obey when threat
1	2	-.01	.19	.63	.63	e 1	10. Acts defiant when told to

Finally, a factor sensitivity ratio (Wright & Stone, 2004) was used to determine to what extent the measure is impacted by the secondary dimension. It is calculated by dividing the residual variance eigenvalue units of the first contrast by the variance explained eigenvalue units to generate a ratio of the Rasch dimension that is impacted by the secondary dimension. The factor sensitivity ratio was 0.15, which suggests that 15% of the measure is affected by the unexplained relationships between items.

ODBTA: Model fit. Real world data are not expected to fit Rasch-based models, including RSMs, perfectly, as the models are idealizations (Linacre, 2017). In fact, it is expected that global fit statistics, such as chi-square, will show significant misfit to the model. Therefore, a variety of Rasch fit statistics were utilized to evaluate whether the data conformed to the model enough to measure oppositional defiant behaviors meaningfully. Fit statistics help to assess whether the data deviate from model expectations sufficiently to distort measurement significantly. Evaluation of the fit of the

data to the RSM included examination of item polarity, item-category measures, and item and person fit mean-square statistics, i.e., infit and outfit.

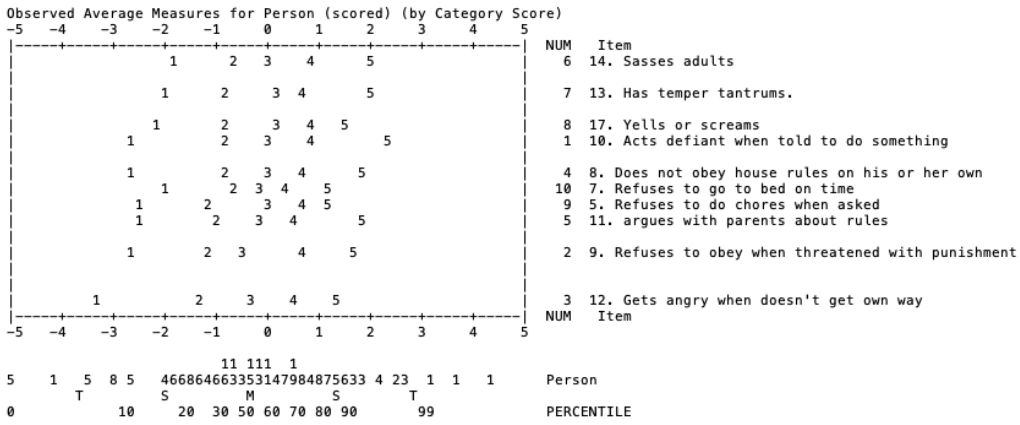
Item-polarity. Item polarity shows the degree to which items align with the latent variable (Bond & Fox, 2015). Polarity is estimated using point-measure correlations as reported by WINSTEPS and is related to the fundamental assumption that higher ability aligns with higher ratings on items and vice versa (Linacre, 2017). Positive point-measure correlations suggest that higher response options align with higher levels of the intended construct. Negative point-measure correlations suggest that the item's orientation does not align with the intended construct. Negative point-measure correlations may be caused by reverse-scoring, guessing, entry errors, or randomness in the data. All 10 of the ODBTA items were found to have positive point-measure correlations, suggesting that all the items aligned with the underlying construct as expected.

Average person measures plot. The category functioning for each item was further assessed by visual inspection of the observed average measures for each scored category plot (Figure 4) in order to confirm that higher trait levels resulted in endorsing a higher category (Linacre, 2017). Figure 4 shows that for each item of the ODBTA scale, the average measures of the sample support the assertion that endorsing a higher category aligns with higher severity of oppositional defiant behavior. This is evidenced by the ordering of the categories for each item in the plot. Additionally, Figure 4 also shows the item hierarchy for the ODBTA items. The item hierarchy helps to define the latent variable that is measured by the items. In this case, the latent variable is the severity of oppositional defiant behavior toward adults. The items are listed from the most difficult to endorse, i.e., sasses adults, to the easiest to endorse, i.e., gets angry when doesn't get

own way. Furthermore, the person measures at the bottom of Figure 4 show the distribution of the sample on the latent trait, where *M* is the location of the average person measure. For the ODBTA scale, the average person measure is just below the local origin, which is indicated by "0" on the measurement scale. Lastly, Figure 4 shows that the category measures for each item are within the majority of the range of the sample's person measures.

Figure 4

Observed Average Measures Plot of the ODBTA Items



Item fit. Item fit was evaluated using infit and outfit mean-square statistics. Infit is based on the chi-square statistic, where each observation is weighted by its statistical information, i.e., model variance (Linacre, 2017). For the infit and outfit mean-square fit statistics, values close to one and within the range of 0.5 to 2.0 suggest adequate fit to the RSM (Linacre, 2017). High outfit values may be the result of random responses or outliers. High infit values are influenced by unexpected response patterns and are more likely to be a threat to measurement. Alternatively, low mean-square values suggest that the observations may be too predictable and overfitting.

Item fit is the extent to which the items function differently from the expectations of the measurement model. Item fit mean-square statistics for the ODBTA factor can be found in Table 12. Infit mean-square values ranged from 0.63 to 1.78 (M = 1.00, SD = 0.30) and outfit mean-square values ranged from 0.63 to 1.84. (M = 1.02, SD = 0.33). These values are within the expected range and do not suggest item misfit.

Table 12

ODBTA Item Fit Statistics

Item	Infit Mean-square Statistic	Outfit Mean-square Statistic
7.	1.78	1.84
5.	1.12	1.24
14.	1.13	1.07
17.	1.07	1.05
13.	1.01	1.06
8.	0.84	0.94
9.	0.83	0.80
11.	0.83	0.83
12.	0.74	0.72
10.	0.63	0.63

Person fit. Person fit is the extent to which responses differ from the expectations of the measurement model (Bond & Fox, 2015; Linacre, 2017). Similar to item fit, infit and outfit mean-square statistics are used to identify misfitting persons. Mean-square values larger than 2.0 suggest that the person responded in an unexpected manner (Bond & Fox, 2015). Additionally, mean-square values larger than 2.0 can distort measurement but can be caused by a few unexpected responses (Linacre, 2017). Unexpected responses are considered to be degrading to measurement, i.e., to estimates of person and item measures, when there is a large number of person fit mean-squares outside of the expected range. A summary of the person fit statistics for the 216 non-extreme persons can be found in Table 13. The infit mean-square values for the ODBTA factor ranged

from 0.05 to 3.59 (*M* = 1.02, *SD* = 0.65), and outfit mean-square values ranged from 0.05 to 3.72 (*M* = 1.02, *SD* = 0.66). Five persons, or 2.3%, responded in an extreme manner.

Table 13

ODBTA Person Fit Statistics Summary

	Infit MNSQ	Outfit MNSQ
Mean	1.02	1.02
Standard Deviation	0.65	0.66
Maximum	3.59	3.72
Minimum	0.05	0.05

ODBTA: Differential item functioning. Differential item functioning (DIF) was investigated in order to assess the degree to which the items of the ODBTA scale function similarly across Hispanic and non-Hispanic, i.e., European Americans, African Americans, "other," groups. The presence of significant DIF would suggest that the probability of endorsing an item is different between groups when the person measure, i.e., severity of oppositional defiant behavior toward adults, is constant (Furr, 2018). DIF may suggest item bias and may indicate the presence of a secondary trait. Significance testing, i.e., pair-wise comparisons of the DIF measures between Hispanics and each other ethnic group, as well as examination of the DIF contrast, i.e., the difference in DIF between each of the comparisons (i.e., Hispanics compared individually with each other ethnic group), was used to assess DIF (Furr, 2018; Linacre, 2017). |DIF contrast| ≥ 0.5 logits (Linacre, 2017) was considered to be substantive.

The total sample included 221 extreme (n = 5) and non-extreme (n = 216) persons. Extreme persons are uninformative to DIF analysis, as they do not contribute to the estimation of item difficulty (Linacre, 2017). Therefore, DIF analysis was based on the 216 non-extreme persons. The reference group for the DIF analysis was the Hispanic

group (n = 95), and the focal groups compared to the reference group were the European American group (n = 89), African American group (n = 26), and "other" group (n = 6). Given the notably small sample size of the "other" and African American groups, DIF results relating to those groups were considered exploratory rather than decisive (Linacre, 2017).

When making multiple comparisons, the chance of committing type one errors, i.e., incorrectly rejecting the null hypothesis, increases. In order to decrease the chance of making type one errors and observing significance due to chance when many comparisons are being made, a Bonferroni correction was used (Bonferroni, 1936). The Bonferroni correction is a method used in multiple hypothesis testing that aids in controlling the occurrence of false positives (Abdi, 2007). The Bonferroni correction accounts for the increase in risk of type one errors by modifying the alpha level to account for the multiple comparisons being made. To make this correction, the alpha value was divided by the number of pair-wise comparisons, i.e. 30. The correction resulted in an alpha value of 0.0016.

Results of the DIF analysis between the Hispanic and European American group, the Hispanic and African American group, and the Hispanic and "other" group, are shown in Table 14. The pairwise comparisons between the Hispanic group and the European American, the African American, and the "other" groups did not result in statistically significant DIF for the items of the ODBTA scale. However, several items were found to have considerable values for DIF contrast, i.e., $|\text{DIF contrast}| \geq 0.5$ logits, without statistical significance. Specifically, item 11 was 0.65 logits more difficult for African Americans (DIF Measure = 0.43) than for Hispanics (DIF Measure = -0.22), item

13 was 0.77 logits more difficult for persons in the "other" group (DIF Measure = 0.98) than for Hispanics (DIF Measure = 0.20), and item five was 0.55 logits more difficult for Hispanics (DIF Measure = -0.05) than for individuals in the "other" group (DIF Measure= -0.60).

Table 14

DIF for the ODBTA Scale

Item	Hispanic Group DIF	Comparison Group	Comparison Group DIF	DIF Contrast	Mantel χ^2	Probability
10.	0.27	EA	0.19	0.08	0.0092	0.9234
10.	0.27	AA	-0.04	0.31	1.1732	0.2787
10.	0.27	Other	-0.10	0.37	0.0110	0.9166
9.	-0.41	EA	-0.15	-0.26	5.0113	0.0252
9.	-0.41	AA	-0.39	-0.02	0.0513	0.8208
9.	-0.41	Other	-0.10	-0.31	0.3067	0.5797
12.	-0.51	EA	-0.56	0.05	0.2149	0.6429
12.	-0.51	AA	-0.73	0.23	1.2751	0.2588
12.	-0.51	Other	-0.60	0.10	2.3787	0.1230
8.	-0.03	EA	0.09	-0.12	0.0413	0.8390
8.	-0.03	AA	-0.04	0.01	0.0625	0.8026
8.	-0.03	Other	-0.10	0.07	0.5390	0.4629
11.	-0.22	EA	-0.28	0.05	0.0787	0.7791
11.	-0.22	AA	0.43	**-0.65**	9.0984	0.0026
11.	-0.22	Other	-0.60	0.38	1.4935	0.2217
14.	0.62	EA	0.26	0.36	0.4481	0.5033
14.	0.62	AA	0.37	0.25	0.0124	0.9113
14.	0.62	Other	0.69	-0.06	0.4874	0.4851
13.	0.2	EA	0.4	-0.20	0.3641	0.5462
13.	0.2	AA	0.49	-0.29	0.0073	0.9320
13.	0.2	Other	0.98	**-0.77**	1.3546	0.2445
17.	0.25	EA	0.20	0.05	0.0302	0.8620
17.	0.25	AA	0.02	0.23	1.6783	0.1951
17.	0.25	Other	0.41	-0.16	0.0390	0.8434
5.	-0.05	EA	-0.17	0.12	1.7506	0.1858
5.	-0.05	AA	-0.04	-0.01	0.2990	0.5845
5.	-0.05	Other	-0.60	**0.55**	0.0879	0.7669
7.	-0.11	EA	0.02	-0.13	0.0068	0.9341
7.	-0.11	AA	-0.04	-0.07	0.3061	0.5801
7.	-0.11	Other	0.15	-0.26	0.3536	0.5521

Note: *|DIF contrast| ≥ 0.5 logits in bold.* EA = European American and AA= African American.
**p-value* < .0016

CPB: Rating scale functioning. The eight IS items of the Conduct Problem Behavior (CPB) scale, Table 15, were used for this analysis. Each item shared the same seven-point rating scale structure. Response options ranged from "never" (i.e., one); "seldom" (i.e., two or three); "sometimes" (i.e., four); "often" (i.e., five or six); to "always" (i.e., seven). As noted above, while a PCM was initially planned for this analysis, the RSM was ultimately selected given the shared rating scale structure of the items.

Table 15

Items of the CPB ECBI Scale

19. Destroys toys and other objects
21. Steals
22. Lies
23. Teases or provokes other children
24. Verbally fights with friends his or her own age
25. Verbally fights with sisters and brothers
26. Physically fights with friends his or her own age
27. Physically fights with sisters and brothers

A specification of the RSM is the assumption of monotonicity. This indicates that it is expected that as trait level increases, e.g. severity of problematic conduct behavior, so does the probability of endorsing a higher response category, e.g., selecting a higher item severity such as "always" (Embretson & Reise, 2000; Linacre, 2017). Violation of this specification would suggest that the categories may be disordered and not functioning in accordance with the model.

Monotonicity for the CPB rating scale was assessed by examining the ordering, or disordering, of observed averages and the ordering, or disordering, of Andrich thresholds (Andrich, 2006). The observed averages for a category are the average abilities of the

respondents who endorsed that category (Linacre, 2017). In other words, observed averages are the average ability, or trait, levels needed for a respondent in this sample to endorse a certain response option. Similar to Andrich thresholds, a specification of the RSM is that observed averages will increase as response options advance (Embreston & Reise, 2000; Linacre, 2017). Ordered observed averages are an indication that trait level advances as categories advance. An Andrich threshold, or step difficulty, is the point at which there is a 50-50 probability of endorsing adjacent response options and is indicated by the point at which response probability curves intersect for adjacent response options (Andrich, 2006; Linacre, 2017). For polytomous items, there are k -1 thresholds, where k is the number of category options. It is expected that thresholds advance as trait level increases. Additionally, when thresholds are ordered, each category is most probable at some point along the scale.

In addition to monotonicity, category mean-square fit statistics, outfit and infit, were evaluated to assess category usage (Linacre, 1995). Mean-square statistic values near 1.0 indicate appropriate category usage (Linacre, 2017). Values greater than 2.0 indicate unpredictability in category usage that can distort measurement, and values less than 0.5 indicate overly predictable usage.

The results in Table 16 show the observed averages, Andrich thresholds, and category mean-square fit statistics for the CPB rating scale items with seven response options. The results show ordered observed averages and category mean-square values within the expected range of 0.5 to 2.0 but with disordered thresholds. Specifically, the threshold between the response options three and four is less than the threshold between

response options one and two, and two and three, and the threshold between the response options four and five is greater than the threshold between response options five and six.

Table 16

CPB Seven-point Rating Scale Functioning

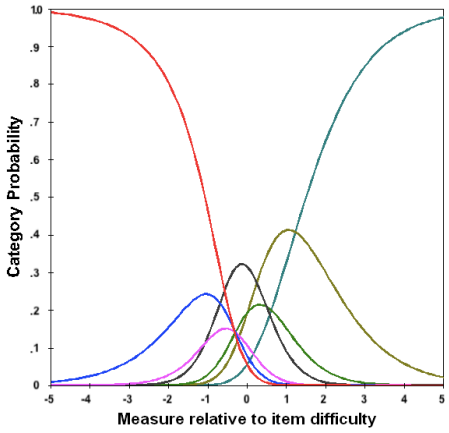
	Response Options						
	1	2	3	4	5	6	7
Andrich Threshold	None	-0.35	-0.29	-1.09	0.49	0.00	1.23
Observed Averages	-1.61	-1.05	-0.78	-0.50	-0.25	-0.04	0.30
Infit MNSQ	1.12	1.03	0.87	.1.03	0.99	1.15	0.91
Outfit MNSQ	1.09	0.73	0.69	0.99	1.02	0.91	0.91

Note: MNSQ = Mean-square statistic.

The disordering of thresholds is depicted by the category probability curve for item 24 in Figure 5. These curves show the probability of observing each category. The Andrich thresholds are the points at which adjacent curves intersect. As each item shares the same rating scale structure for RSMs, each item has the same curve.

Figure 5

Category Probability Curve for Item 24 of the CPB Seven-point Scale



Note: Red = category one probability. Blue = category two probability. Pink = category three probability. Black = category four probability. Green = category five probability. Olive = category six probability. Teal = category seven probability.

When thresholds are found to be disordered, adjustments to the scale can help the rating scale conform to model expectations (Embreston & Reise, 2000; Linacre, 2017). In such cases, collapsing adjacent categories can aid in improving rating scale functioning by addressing disordered thresholds (Embreston & Reise, 2000). Additionally, threshold disordering is often observed when categories correspond to a narrow interval of the latent variable (Linacre, 2017), which can be argued is the case for categories two and three, and five and six of the CPB rating scale. In fact, categories two and three are both labeled "seldom" and categories five and six are both labeled "often." Therefore, combining categories two and three, and categories five and six, and then re-assigning point-scores to the combined categories is reasonable given the limited differentiation in their descriptions.

The responses for the CPB scale items were recoded using WINSTEPS from a seven-point rating scale to a five-point rating scale. The resulting response options for these items were "never" (i.e., one); "seldom" (i.e., two being combined categories two and three); "sometimes" (i.e., three); "often" (i.e., four being combined categories five and six); to "always" (i.e., five). The data were then re-evaluated using the RSM in order to assess the rating scale functioning. Specifically, observed averages, Andrich thresholds, and category mean-square fit statistics were examined for the collapsed five-point rating scale. The results for the items with a five-point rating scale are presented in Table 17. The observed averages are ordered, and the category mean-square fit statistics are within the expected range of 0.5 to 2.0. Additionally, the Andrich thresholds are ordered. These results show that for the CPB items the five-point rating scale is functioning in accordance with the RSM specifications.

Table 17

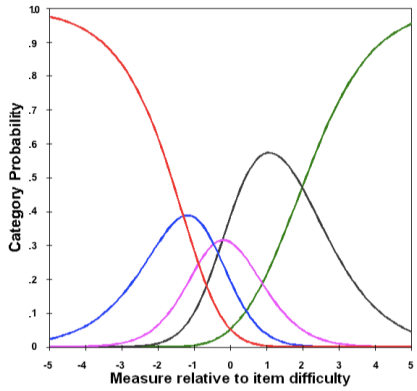
CPB Five-point Rating Scale Functioning

	Response Options				
	1	2	3	4	5
Andrich Threshold	None	-1.30	-0.46	-0.22	1.98
Observed Averages	-2.37	-1.39	-0.75	-0.21	0.47
Infit MNSQ	1.13	0.89	1.02	1.08	0.93
Outfit MNSQ	1.10	0.68	1.00	1.15	0.92

Finally, examination of the category probability curve for the CPB items confirms that for the RSM with a five-point rating scale, as trait level increases, so will the probability of endorsing a more advanced category. A category probability curve for item 24 of the CPB scale is shown in Figure 6.

Figure 6

Category Probability Curve for Item 24 of the CPB Five-point Scale



Note: Red = category one probability. Blue = category two probability. Pink = category three probability. Black = category four probability. Green = category five probability.

CPB: Dimensionality. In order to assess the dimensionality of the CPB factor, a PCA of the probability residuals was used (Bond & Fox, 2015; Linacre, 2017). The RSM assumes unidimensionality, which indicates that all the items of the test, or, in this case, of the CPB scale, measure one underlying construct, i.e., problematic conduct behaviors

(Bond & Fox, 2015). It is expected that the Rasch dimension, i.e., the person measures and the item measures, will explain the majority of the variance in the data. Therefore, in order to assess dimensionality, the raw variance explained by measures, or the variance that can be explained by the Rasch measures, was evaluated. At least 40% of the variance explained by measures was tentatively used to evaluate unidimensionality (Linacre, 2017). The results of the PCA of probability of residuals for the CPB factor are presented in Table 18. The raw explained variance was 47.8%, and above the established guideline of 40%, suggesting that the CPB scale is unidimensional enough for meaningful measurement of conduct problem behaviors.

Table 18

Results of the PCA of Residuals of the CPB Scale

	Eigenvalue	Percentage
Total Raw Variance Explained by Measures	7.32	47.8%
Raw Variance Explained by Persons	2.44	15.9%
Raw Variance Explained by Items	4.89	31.9%
Total Unexplained Variance	8.00	52.2%
Unexplained Variance in 1st Contrast	2.19	14.3%
Unexplained Variance in 2nd Contrast	1.64	10.7%
Unexplained Variance in 3rd Contrast	1.26	8.20%
Unexplained Variance in 4th Contrast	1.01	6.60%
Unexplained Variance in 5th Contrast	0.80	10.1%

In order to assess further the dimensionality of the CPB factor, the raw unexplained variance was also used to determine whether another meaningful dimension, after the Rasch dimension had been accounted for, explained a significant amount of the residual variance (Bond & Fox, 2015; Linacre, 2017). The presence of a meaningful dimension would suggest multidimensionality in the data. The unexplained variance was explored using a PCA of the residual variance after the Rasch dimension has been

accounted for. If the data were unidimensional, the components, or factors, identified by the PCA would be expected to be at “noise” level (Linacre, 1998, 2017; Wright, 1996). Therefore, less than 15% of unexplained variance in the first contrast along with an eigenvalue less than two were tentative guidelines for assessing unidimensionality. The unexplained variance accounted for by the first contrast was 14.3% with an eigenvalue of 2.19 (Table 18).

Since the eigenvalue for the first contrast was slightly above the expected value of two, examination of the content or the wording of the items found to share residual variance that differed from the other items aided in further clarifying the dimensionality of the CPB factor. Figure 7 shows the items that were found to share residual variance that may indicate a second dimension.

The summary of residual loadings (Figure 7) shows that items at the top of the table, i.e., items 25 and 27, share content related to siblings while the items toward the bottom of the table do not. However, aside from this, the items all share similar content related to problematic conduct behavior. Given that the CPB scale is intended to measure severity of problematic conduct behaviors and that all of the items reflect general problematic conduct behaviors, it would neither be conceptually sound nor improve measurement to split the items into two dimensions. Overall, the items reflect general problematic conduct behaviors, suggesting that the items are unidimensional enough for adequate measurement of problematic conduct behaviors.

Figure 7

Contrast Plot of the CPB Items' Residual Loadings

CON TRA	CL US	LOADING	INFIT MEASURE	OUTFIT MNSQ	ENTRY MNSQ	NUMBER	Item
1	1	.88	-1.01	1.22	1.17	8	8 25. Verbally fights with siste
1	1	.81	-.17	1.02	.89	6	6 27. Physically fights with sis
1	3	-.51	-.05	.83	.82	1	1 24. Verbally fights with frien
1	3	-.39	.28	1.13	1.09	4	4 19. Destroys toys and other ob
1	3	-.37	.96	1.06	.94	2	2 26. Physically fights with fri
1	3	-.28	-.84	.96	1.01	5	5 22. Lies
1	3	-.26	-.24	.97	.86	3	3 23. Teases or provokes other c
1	3	-.24	1.05	1.28	.99	7	7 21. Steals

Finally, a factor sensitivity ratio (Wright & Stone, 2004) was used to determine to what extent the measure is impacted by the secondary dimension. It is calculated by dividing the residual variance eigenvalue units of the first contrast by the variance explained eigenvalue units to generate a ratio of the Rasch dimension that is impacted by the secondary dimension. The factor sensitivity ratio was 0.29, which suggests that 29% of the measure is affected by the unexplained relationships between items.

CPB: Model fit. Real world data are not expected to fit Rasch-based models, including RSMs, perfectly as they are idealizations (Linacre, 2017). In fact, it is expected that global fit statistics, such as the chi-square, will show significant misfit to the model. Therefore, a variety of Rasch fit statistics were utilized to evaluate whether the data conformed to the model enough to measure problematic conduct behaviors meaningfully. Fit statistics help to assess whether the data deviate from model expectations sufficiently to distort measurement significantly. Evaluation of the fit of the data to the RSM included examination of item polarity, item-category measures, and item and person fit mean-square statistics, i.e., infit and outfit.

Item-polarity. Item polarity shows the degree to which items are aligned with the latent variable (Bond & Fox, 2015). Polarity is assessed using point-measure correlations as reported by WINSTEPS and is related to the fundamental assumption that higher ability aligns with higher ratings on items, and vice versa (Linacre, 2017). Positive point-measure correlations suggest that the item measures the intended construct. Negative point-measure correlations suggest that the item's orientation does not align with the intended construct. Negative point-measure correlations may be caused by reverse-scoring, guessing, entry errors, or randomness in the data. All eight of the CPB items were found to have positive point-measure correlations, suggesting that all the items aligned with the underlying construct as expected.

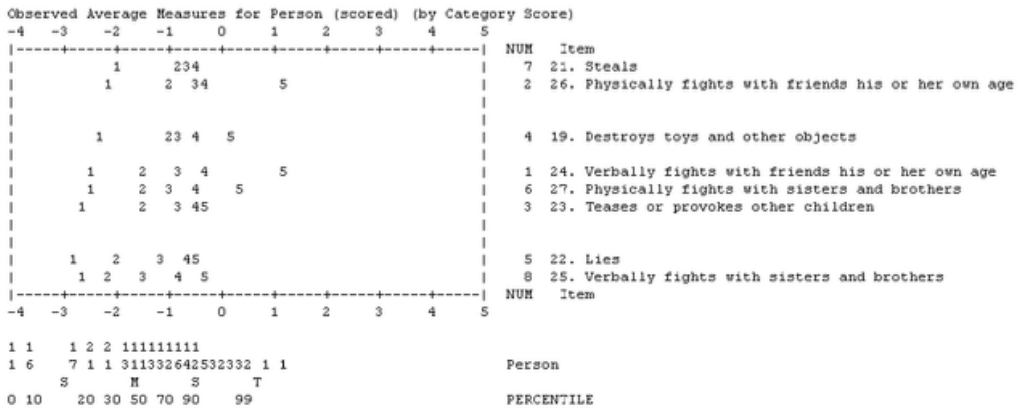
Average person measures plot. The category functioning for each item was assessed by visual inspection of the item-category measures plot (Figure 8) in order to confirm that higher trait levels resulted in endorsing a higher category (Linacre, 2017). Figure 8 shows that for each item of the CPB scale, the average measures for each category support the assertion that endorsing a higher category aligns with higher severity of problematic conduct behaviors. This is evidenced by the ordering of the category numbers from left to right for each item. Figure 8 also shows that category five for item 21 of the CPB scale is not depicted on the plot, indicating that response option five, i.e., "always" was not endorsed for item five.

Figure 8 also shows the item hierarchy for the CPB items. The item hierarchy helps to define the latent variable that is measured by the items. In this case, the latent variable is the severity of problematic conduct behaviors. The items are listed from the most difficult to endorse, i.e., steals, to the easiest to endorse, i.e., verbally fights with

sisters and brothers. Furthermore, the person measures at the bottom of Figure 8 show the distribution of the sample on the latent trait, where *M* is the location of the average person measure. For the CPB scale, the average person measure is within one to two logits below the local origin, which is indicated by "0" on the measurement scale. Lastly, Figure 8 shows that the plotted item measures fall within the range of the sample's person measures.

Figure 8

Observed Average Measures Plot of the CPB Items



Item fit. Item fit was evaluated using infit and outfit mean-square statistics. Infit is based on the chi-square statistic, where each observation is weighted by its statistical information, i.e., model variance (Linacre, 2017). For the infit and outfit mean-square fit statistics, values close to one and within the range of 0.5 to 2.0 suggest adequate fit to the RSM (Linacre, 2017). High outfit values may be the result of random responses or outliers. High infit values are influenced by inlier response patterns and are more likely to be a threat to measurement. Alternatively, low mean-square values suggest that the observations may be too predictable and overfitting.

Item fit is the extent to which the items function differently from the expectations of the measurement model. Item fit mean-square statistics for the CPB scale can be found in Table 19. Infit mean-square values ranged from 1.28 to 0.83 (M = 1.06, SD = 0.14) and outfit mean-square values ranged from 1.17 to 0.82 (M = 0.97, SD = 0.11). These values are within the expected range and do not suggest item misfit.

Table 19

CPB Item Fit Statistics

Item	Infit Mean-square Statistic	Outfit Mean-square Statistic
7.	1.28	0.99
8.	1.22	1.17
4.	1.13	1.09
2.	1.06	0.94
6.	1.02	0.89
5.	0.96	1.01
3.	0.97	0.86
1.	0.83	0.82

Person fit. Person fit is the extent to which responses differ from the expectations of the measurement model (Bond & Fox, 2015; Linacre, 2017). Similar to item fit, infit and outfit mean-square statistics are used to identify misfitting persons. Mean-square values larger than 2.0 suggest that the person responded in an unexpected manner (Bond & Fox, 2017). Furthermore, large mean-square values can distort measurement but can be caused by a few unexpected responses (Linacre, 2017). Unexpected responses are considered to be degrading to measurement, i.e., estimates of person and item measures, when there is a large number of person fit mean-squares outside of the expected range. A summary of the person fit statistics can be found in Table 20. The infit mean-square values for the CPB scale had a mean of 0.99 and ranged from 3.12 to 0.18 (SD = 0.61)

and outfit mean-square values had a mean of 0.97 and ranged from 4.64 to 0.22 (*SD* = 0.73). Eleven persons, or 5% of the sample, responded in an extreme manner.

Table 20

CPB Person Fit Statistics Summary

	Infit MNSQ	Outfit MNSQ
Mean	0.99	0.97
Standard Deviation	0.61	0.73
Maximum	3.12	4.64
Minimum	0.18	0.22

CPB: DIF. DIF was investigated in order to assess the degree to which the items of the CPB scale function similarly across Hispanic and non-Hispanic (i.e., European Americans, African Americans, and "other") groups. The presence of significant DIF would suggest that the probability of endorsing an item is different between groups when the person measure, i.e., severity of problematic conduct behavior, is constant (Furr, 2018). Furthermore, significant DIF may suggest item bias and/or may indicate the presence of a secondary trait. Significance testing, i.e., pair-wise comparison of the DIF measures between ethnic groups, as well as examination of the DIF contrast, i.e., the difference in DIF between the two groups, was used to assess DIF (Linacre, 2017). |DIF contrast| ≥ 0.5 logits (Linacre, 2017) was considered to be substantive.

The total sample included 221 extreme (n = 11) and non-extreme (n = 210) persons. Extreme persons are uninformative to DIF analysis, as they do not contribute to the estimations of item difficulties (Linacre, 2017). Therefore, DIF analysis was based on the 210 non-extreme persons. The reference group for the DIF analysis was the Hispanic group (n = 93), and the focal groups compared to the reference group were the European American group (n = 86), African American group (n = 24), and "other" group (n = 7).

Given the notably small sample size of the "other" (n =7) and African American (n = 24) groups, DIF results relating to those groups were considered exploratory rather than decisive (Linacre, 2017).

When making multiple comparisons, the chance of committing type one errors, i.e., incorrectly rejecting the null hypothesis, increases. In order to decrease the chance of type one errors and observing significance due to chance when many comparisons are being made, a Bonferroni correction was used (Bonferroni, 1936). The Bonferroni correction is a method used in multiple hypothesis testing that aids in controlling the occurrence of false positives (Abdi, 2007). The Bonferroni correction accounts for the increase in risk of type one errors by modifying the alpha level, i.e., 0.05, to account for the multiple comparisons being made. To make this correction, the alpha value was divided by the number of pair-wise comparisons. The correction resulted in an alpha value of 0.002.

Results of the DIF analysis between Hispanic and European American groups, Hispanic and African American groups, and Hispanic and "other" groups, are shown in Table 21. The pairwise comparisons of DIF measures between the Hispanic group and the European American, the African American, and the "other" groups, did not indicate statistically significant DIF, i.e., $p < 0.002$, for the items of the CPB scale. Although, six items were found to have considerable values for DIF contrast, i.e., $|\text{DIF contrast}| \geq 0.5$ logits, although without statistical significance. Specifically, three items were more difficult for individuals in the "other" group than for Hispanics. Item 26 was 1.06 logits more difficult for individuals in the "other" group (DIF Measure = 2.07) than for Hispanics (DIF Measure = 1.01), item 23 was 1.01 logits more difficult for individuals in

the “other” group (DIF Measure = 0.85) than for Hispanics (DIF Measure = -0.16), and item 27 was 2.34 logits more difficult for individuals in the “other” group (DIF Measure = 2.05) than for Hispanics (DIF Measure = -0.29). Additionally, three items were found to be more difficult for Hispanics than for individuals in the “other” group. Item 19 was 1.15 logits more difficult for Hispanics (DIF Measure = 0.54) than for individuals in the “other” group (DIF Measure = -0.61), item 22 was 0.76 logits more difficult for Hispanics (DIF Measure = -0.86) than for individuals in the “other” group (DIF Measure = -1.62), and item 21 was 0.5 logits more difficult for Hispanics (DIF Measure = 1.35) than for individuals in the “other” group (DIF Measure = 0.85). Finally, item 21 was also 0.8 logits more difficult for Hispanics (DIF Measure = 1.35) than for African Americans (DIF Measure = 0.55).

Table 21

DIF for the CPB Scale

Item	Hispanic Group DIF	Comparison Group	Comparison Group DIF	DIF Contrast	χ^2	Probability
24.	-0.14	EA	-0.05	-0.10	0.7472	0.3874
24.	-0.14	AA	0.3	-0.45	3.0606	0.0802
24.	-0.14	Other	0.14	-0.28	0.0571	0.8112
26.	1.01	EA	0.85	0.16	0.0253	0.8736
26.	1.01	AA	1.07	-0.06	0.2267	0.6340
26.	1.01	Other	2.07	**-1.06**	1.0251	0.3113
23.	-0.16	EA	-0.41	0.25	2.3454	0.1257
23.	-0.16	AA	0.08	-0.24	0.0033	0.9541
23.	-0.16	Other	0.85	**-1.01**	0.3881	0.5333
19.	0.54	EA	0.06	0.48	3.1124	0.0777
19.	0.54	AA	0.46	0.08	0.1831	0.6687
19.	0.54	Other	-0.61	**1.15**	4.1238	0.0423
22.	-0.86	EA	-0.63	-0.23	0.8237	0.3641
22.	-0.86	AA	-1.32	0.46	5.9743	0.0145
22.	-0.86	Other	-1.62	**0.76**	0.3314	0.5648
27.	-0.29	EA	-0.11	-0.18	0.7480	0.3871
27.	-0.29	AA	-0.13	-0.16	1.2157	0.2702
27.	-0.29	Other	2.05	**-2.34**	1.2134	0.2707
21.	1.35	EA	0.96	0.39	2.0708	0.1501
21.	1.35	AA	0.55	**0.80**	6.0262	0.0141
21.	1.35	Other	0.85	**0.50**	4.0000	0.0455
25.	-1.17	EA	-0.85	-0.32	1.9046	0.1676
25.	-1.17	AA	-1.01	-0.16	3.4961	0.0615
25.	-1.17	Other	-0.86	-0.31	0.4952	0.4816

Note: |*DIF contrast*| ≥ *0.5 logits in bold.* EA = European American, AA= African American.
**p-value* < .002

IB: Rating scale functioning. The four IS items of the Inattentive Behavior (IB) scale, Table 22, were used for this analysis. Each item shared the same seven-point rating scale structure. Response options ranged from “never” (i.e., one); “seldom” (i.e., two or three); “sometimes” (i.e., four); “often” (i.e., five or six); to “always” (i.e., seven). As noted above, while a PCM was initially planned for this analysis, the RSM was ultimately selected, given the shared rating scale structure of the items.

Table 22

Items of the IB ECBI Scale

IB Scale
31. Has short attention span
30. Is easily distracted
34. Has difficulty concentrating on one thing
32. Fails to finish tasks or projects

A specification of the RSM is the assumption of monotonicity. This reflects the expectation that as trait level, e.g., severity of problematic behavior, increases, so does the probability of endorsing a higher response category, e.g., selecting a higher item severity such as "always" (Embretson & Reise, 2000; Linacre, 2017). Violation of this specification would suggest that the categories may be disordered and are not functioning in accordance with the model.

Monotonicity for the IB rating scale was assessed by examining the ordering or disordering of observed averages and the ordering or disordering of Andrich thresholds (Andrich, 2006). The observed averages for a category reflect the average trait levels as rated by the respondents who endorsed that category (Linacre, 2017). In other words, observed averages reflect the average ability, or trait, levels needed for a respondent in this sample to endorse a certain response option. Similar to Andrich thresholds, a specification of the RSM is that observed averages will increase as response options advance (Embreston & Reise, 2000; Linacre, 2017). Ordered observed averages are an indication that trait level advances as categories advance. An Andrich threshold, or step difficulty, is the point at which there is a 50-50 probability of endorsing adjacent response options and is indicated by the point at which response probability curves intersect for adjacent response options (Andrich, 2006; Linacre, 2017). For polytomous

items, there are *k* -1 thresholds, where *k* is the number of category options. It is expected that thresholds advance as trait level increases. Additionally, when thresholds are ordered, each category is most probable at some point along the scale.

In addition to monotonicity, category mean-square fit statistics, outfit and infit, were evaluated to assess category usage (Linacre, 1995). Mean-square statistic values near 1.0 indicate appropriate category usage (Linacre, 2017). Values greater than 2.0 indicate unpredictability in category usage that can distort measurement, and values less than 0.5 indicate overly predictable usage.

The results in Table 23 show the observed averages, Andrich thresholds, and category mean-square fit statistics for the IB rating scale items with seven response options. The results show disordered thresholds and large mean-square values for category one. This suggests that the scale is not functioning as expected. Specifically, the threshold between the response options two and three is larger than the threshold between response options three and four. Disordered thresholds may indicate that some categories reflect narrow intervals of the latent variable (Linacre, 2017). Additionally, the category mean-square statistics for category one are larger than 2.0 which suggests unpredictability in category use that can distort measurement.

Table 23

IB Seven-point Rating Scale Functioning

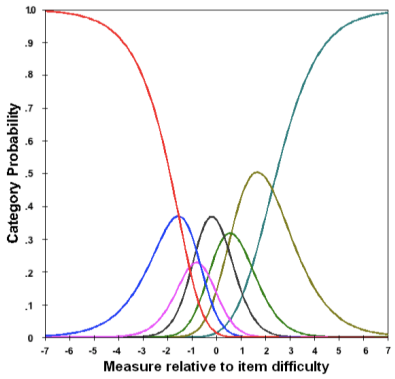
	Response Options						
	1	2	3	4	5	6	7
Andrich Threshold	None	-1.54	-0.67	-0.98	0.32	0.60	2.28
Observed Averages	-2.91	-1.55	-0.81	-0.19	0.56	1.67	3.51
Infit MNSQ	3.04	0.79	0.75	0.61	0.61	0.77	1.13
Outfit MNSQ	3.45	0.91	0.78	0.54	0.70	0.74	1.03

Note: MNSQ = Mean-square statistic.

The disordering of thresholds is depicted by the category probability curve for item 31 in Figure 9. These curves show the probability of endorsing each category. The Andrich thresholds are the points at which adjacent curves intersect. Since each item shares the same rating scale structure for RSMs, each item has the same curve.

Figure 9

Category Probability Curve for Item 31 of the IB Seven-point Scale



Note: Red = category one probability. Blue = category two probability. Pink = category three probability. Black = category four probability. Green = category five probability. Olive = category six probability. Teal = category seven probability.

When thresholds are found to be disordered, adjustments to the scale are needed to help the rating scale conform to model expectations (Embreston & Reise, 2000; Linacre, 2017). In such cases, collapsing adjacent categories can aid in improving rating scale functioning by addressing disordered thresholds (Embreston & Reise, 2000). Additionally, threshold disordering is often observed when categories correspond to a narrow interval of the latent variable (Linacre, 2017), which can be argued is the case for categories two and three as well as five and six of the IB rating scale. In fact, categories two and three are both labeled “seldom” and categories five and six are both labeled “often.” Therefore, combining categories two and three as well as categories five and six,

and then re-assigning point-scores to the combined categories is reasonable, given the limited differentiation in their descriptions.

The responses for the IB items were recoded using WINSTEPS from a seven-point rating scale to a five-point rating scale. The resulting response options for these items were “never” (i.e., one); “seldom” (i.e., two being combined responses two and three); “sometimes” (i.e., three); “often” (i.e., four being combined responses five and six); to “always” (i.e., five). The data were then re-evaluated using the RSM in order to assess the rating scale functioning. Specifically, observed averages, Andrich thresholds, and category mean-square fit statistics were examined for the collapsed five-point rating scale.

The results for the IB items with a five-point rating scale are presented in Table 24. The observed averages are ordered, the Andrich thresholds are ordered, and the majority of the mean-square fit statistics are within the expected range, i.e., 0.5 to 2.0. However, the infit mean-square statistic for category one (infit MNSQ = 2.23) is somewhat larger than the expected value of 2.0. While this suggests unpredictability in category usage, it is important to consider the possible impact of the ECBI scoring instructions on the category structure indices. The scoring procedures for the ECBI direct examiners to score unanswered items as “never,” i.e., category one. This is problematic for category functioning analyses, as it is unclear whether endorsing category one was the result of a true “never” response or if other reasons led to the endorsement of category one, such as mistakenly skipped items or purposely skipped items due to non-applicability. In general, scoring rules such as these often produce misfit (Linacre, 2017). Notably, despite the higher than expected infit MNSQ for category one, the category

observed averages were not impacted by unpredictable response patterns to an extent that would cause category disordering. Therefore, these results show that for the IB items the five-point rating scale is functioning in a manner that would be productive for measurement.

Table 24

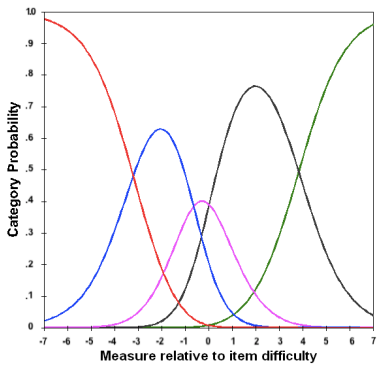
IB Five-point Rating Scale Functioning

	Response Options				
	1	2	3	4	5
Andrich Threshold	None	-3.21	-0.64	0.02	3.83
Observed Averages	-2.42	-1.38	-0.02	1.99	4.18
Infit MNSQ	2.23	1.06	0.74	0.73	1.08
Outfit MNSQ	1.99	1.32	0.68	0.71	0.96

Finally, examination of the category probability curves for the IB items confirms that for the RSM with a five-point rating scale, as trait level increases, so will the probability of endorsing a more advanced category. A category probability curve for item 31 of the IB scale is shown in Figure 10.

Figure 10

Category Probability Curve for Item 31 of the IB Five-point Scale



Note: Red = category one probability. Blue = category two probability. Pink = category three probability. Black = category four probability. Green = category five probability.

IB: Dimensionality. In order to assess the dimensionality of the IB scale, a PCA of the probability residuals was used (Bond & Fox, 2015; Linacre, 2017). The RSM assumes unidimensionality, which indicates that all the items of the test, or, in this case, of the IB scale, measure one underlying construct, i.e., inattentive behaviors (Bond & Fox, 2015). It is expected that the Rasch dimension, i.e., the person measures and the item measures, will explain the majority of the variance in the data. Therefore, in order to assess dimensionality, the raw variance explained by measures, or the variance that can be explained by the Rasch measures, was evaluated. At least 40% of the variance explained by measures was tentatively used to evaluate unidimensionality (Linacre, 2017). The results of the PCA of probability of residuals for the IB scale are presented in Table 25. The raw explained variance was 69.0% which is above the established guideline of 40%, suggesting that the IB scale is unidimensional enough to meaningfully measure inattentive behaviors.

Table 25

Results of the PCA of Residuals of the IB Scale

	Eigenvalue	Percentage
Total Raw Variance Explained by Measures	8.190	69.0%
Raw Variance Explained by Persons	6.710	51.9%
Raw Variance Explained by Items	2.210	17.1%
Total Unexplained Variance	4.000	31.0%
Unexplained Variance in 1st Contrast	1.590	12.4%
Unexplained Variance in 2nd Contrast	1.380	10.7%
Unexplained Variance in 3rd Contrast	1.020	7.9%
Unexplained Variance in 4th Contrast	0.008	0.1%
Unexplained Variance in 5th Contrast	0.002	0.0%

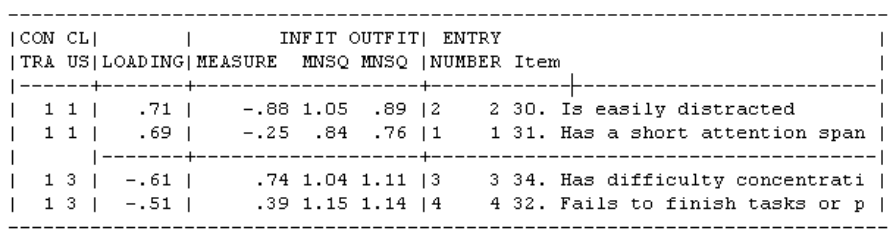
In order to assess further the dimensionality of the IB scale, the raw variance unexplained was also used to determine whether another meaningful dimension, after the

Rasch dimension had been accounted for, explained a significant amount of the residual variance (Bond & Fox, 2015; Linacre, 2017). The presence of a meaningful dimension would suggest multidimensionality in the data. The unexplained variance was explored using a PCA of the residual variance after the Rasch dimension has been accounted for. If the data were unidimensional, the components, or factors, identified by the PCA would be expected to be at "noise" level (Linacre, 1998, 2017; Wright, 1996). Therefore, less than 15% of unexplained variance in the first contrast along with an eigenvalue less than two were tentative guidelines for assessing unidimensionality. The unexplained variance accounted for by the first contrast was 12.4% with an eigenvalue of 1.59 (Table 25).

Examination of the content or the wording of the items found to share residual variance that differed from the other items helped to explain further the dimensionality of the IB scale. Figure 11 shows the contrast plot for the IB items. Comparison of the items at the top of the plot, items 34 and 32, with items at the bottom of the plot, items 31 and 30, suggested that neither dyad of items appeared to share content with each other that would suggest a secondary dimension. Additionally, the dyads did not seem to differ conceptually from each other. Therefore, given that all the IB items are conceptually related to inattentive behaviors, these results suggest that the items are sufficiently unidimensional for the measurement of inattentive behaviors.

Figure 11

Contrast Plot of the IB Items' Residual Loadings



Finally, a factor sensitivity ratio (Wright & Stone, 2004) was used to determine to what extent the measure is impacted by the secondary dimension. It is calculated by dividing the residual unexplained variance eigenvalue units of the first contrast by the variance explained eigenvalue units to generate a ratio of the Rasch dimension that is impacted by the secondary dimension. The factor sensitivity ratio was 0.19 which suggests that 19% of the measure is affected by the unexplained relationships between items.

IB: Model fit. Real world data are not expected to fit Rasch-based models, including RSMs, perfectly as they are idealizations (Linacre, 2017). In fact, it is expected that global fit statistics, such as the chi-square, will show significant misfit to the model. Therefore, a variety of Rasch fit statistics were utilized to evaluate whether the data conform to the model enough to measure inattentive behaviors meaningfully. Fit statistics help to assess whether the data deviate from model expectations sufficiently to distort measurement significantly. Evaluation of the fit of the data to the RSM included examination of item polarity, item-category measures, and item and person fit mean-square statistics, i.e., infit and outfit.

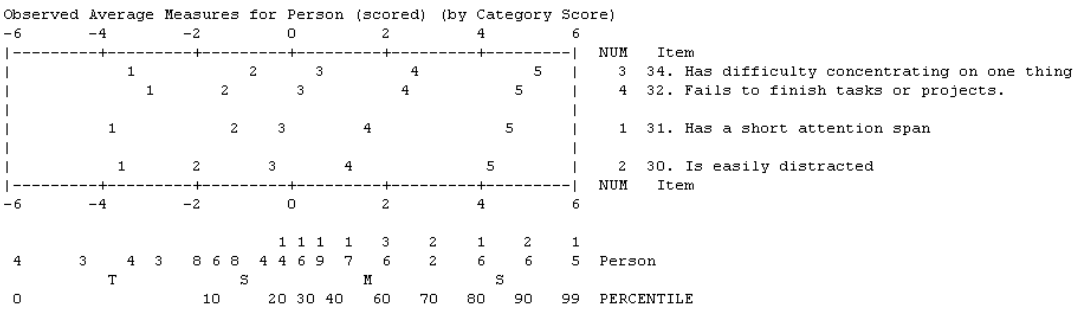
Item-polarity. Item polarity shows the degree to which items are aligned with the latent variable (Bond & Fox, 2015). Polarity is estimated using point-measure correlations as reported by WINSTEPS and is related to the fundamental assumption that higher ability aligns with higher ratings on items and vice versa (Linacre, 2017). Positive point-measure correlations suggest that the item measures the intended construct. Negative point-measure correlations suggest that the item's orientation does not align with the intended construct. Negative point-measure correlations may be caused by

reverse-scoring, guessing, entry errors, or randomness in the data. All four of the IB scale items were found to have positive point-measure correlations, suggesting that all the items aligned with the underlying construct as expected.

Average person measures plot. The category functioning for each item was assessed by visual inspection of the observed average measures for each scored category plot (Figure 11) in order to confirm that higher trait levels resulted in endorsing a higher category (Linacre, 2017). Figure 11 shows that for each item of the IB scale, the average measures for each category support the assertion that endorsing a higher category aligns with higher severity of inattentive behaviors. This is evidenced by the ordering of the category numbers from left to right for each item in the plot. Figure 11 also shows the item hierarchy for the IB scale. The item hierarchy helps to define the latent variable that is being measured by the items. In this case, the latent variable is the severity of inattentive behaviors. The items are listed from the most difficult to endorse, i.e., has difficulty concentrating on one thing, to the easiest to endorse, i.e., is easily distracted. Furthermore, the person measures at the bottom of Figure 11 show the distribution of the sample on the latent trait, where *M* is the location of the average person measure. For the IB scale, the average person measure is less than two logits above the local origin, which is indicated by "0" on the measurement scale. Finally, Figure 11 shows that the plotted category measures for each item fall within the range of the sample's person measures.

Figure 12

Observed Average Measures Plot of the IB Scale



Item fit. Item fit was evaluated using infit and outfit mean-square statistics. Infit is based on the chi-square statistic, where each observation is weighted by its statistical information, i.e., model variance (Linacre, 2017). For the infit and outfit mean-square fit statistics, values close to one and within the range of 0.5 to 2.0 suggest adequate fit to the RSM (Linacre, 2017). High outfit values may be the result of random responses or outliers. High infit values are influenced by inlier response patterns and are more likely to be a threat to measurement. Alternatively, low mean-square values suggest that the observations may be too predictable and overfitting.

Item fit is the extent to which the items function differently from the expectations of the measurement model. Item fit mean-square statistics for the IB scale can be found in Table 26. Infit mean-square values ranged from 0.84 to 1.15 (M = 1.02, SD = 0.11) and outfit mean-square values ranged from 0.76 to 1.14 (M = 0.97, SD = 0.16). These values are within the expected range and do not suggest item misfit.

Table 26

IB Item Fit Statistics

Item	Infit Mean-square Statistic	Outfit Mean-square Statistic
34.	1.04	1.11
32.	1.15	1.14
31.	0.84	0.76
30.	1.05	0.89

Person fit. Person fit is the extent to which responses differ from the expectations of the measurement model (Bond & Fox, 2015; Linacre, 2017). Similar to item fit, infit and outfit mean-square statistics are used to identify misfitting persons. Mean-square values larger than 2.0 suggest that the persons responded in an unexpected manner (Bond & Fox, 2015). Additionally, large mean-square values can distort measurement but can be caused by a few unexpected responses (Linacre, 2017). Unexpected responses are considered to be degrading to measurement, i.e., estimates of person and item measures, when there is a large number of person fit mean-squares outside of the expected range. A summary of the person fit statistics can be found in Table 27. The infit mean-square values for the IB scale ranged from 0.10 to 9.38 (M = 0.99, SD = 1.29) and outfit mean-square values ranged from 0.09 to 9.63 (M = .97, SD = 1.34). Nineteen persons, or 8% of the sample, responded in an extreme manner.

Table 27

IB Person Fit Statistics Summary

	Infit MNSQ	Outfit MNSQ
Mean	0.99	0.97
Standard Deviation	1.29	1.34
Maximum	9.38	9.63
Minimum	0.10	0.09

IB: DIF. DIF was investigated in order to assess the degree to which the items of the IB scale function similarly across Hispanic and non-Hispanic, i.e., European Americans, African Americans, and "other," groups. The presence of significant DIF would suggest that the probability of endorsing an item is different between groups when the person measure, i.e., severity of inattentive behaviors, is constant (Furr, 2018). Furthermore, significant DIF may suggest item bias and/or may indicate the presence of a secondary trait. Significance testing, i.e., pair-wise comparisons of the DIF measures between Hispanics and each other ethnic group, as well as examination of the DIF contrast, i.e., the difference in DIF between each of the comparisons (i.e., Hispanics compared individually with each other ethnic group), was used to assess DIF (Linacre, 2017). |DIF contrast| ≥ 0.5 logits (Linacre, 2017) was considered to be substantive.

The total sample included 221 extreme (n = 19) and non-extreme (n = 202) persons. Extreme persons are uninformative to DIF analysis, as they do not contribute to the estimations of item difficulties (Linacre, 2017). Therefore, DIF analysis was based on the 202 non-extreme persons. The reference group for the DIF analysis was the Hispanic group (n = 88), and the focal groups compared to the reference group were the European American group (n = 83), African American group (n = 24), and "other" group (n = 7). Given the notably small sample size of the "other" and African American groups, DIF results relating to those groups were considered exploratory rather than decisive (Linacre, 2017).

When making multiple comparisons, the chance of committing type one errors, i.e., incorrectly rejecting the null hypothesis, increases. In order to decrease the chance of making type one errors and observing significance due to chance when many

comparisons are being made, a Bonferroni correction was used (Bonferroni, 1936). The Bonferroni correction is a method used in multiple hypothesis testing that aids in controlling the occurrence of false positives (Abdi, 2007). The Bonferroni correction accounts for the increased risk of type one errors by modifying the alpha level, i.e., .05, to account for the multiple comparisons being made. To make this correction, the alpha value was divided by the number of pair-wise comparisons, i.e. 12. The correction resulted in an alpha value of 0.004.

Results of the DIF analysis between Hispanic and European American groups, Hispanic and African American groups, and Hispanic and "other" groups, are shown in Table 28. The pairwise comparisons of DIF measures between the Hispanic group and the European American, the African American, and the "other" groups, did not indicate statistically significant DIF, i.e., $p < 0.004$, for the items of the IB scale. However, item 30 and was found to have a considerable value for DIF contrast, i.e., $|\text{DIF contrast}| \geq 0.5$ logits, without statistical significance. Item 30 was 0.63 logits more difficult for individuals in the "other" group (DIF Measure = -0.49) than for individuals in the Hispanic group (DIF Measure = -1.12).

Table 28

DIF for the IB Scale

Item	Hispanic Group DIF	Comparison Group	Comparison Group DIF	DIF Contrast	Mantel χ^2	Probability
31.	-0.17	EA	-0.32	0.14	0.9028	0.3420
31.	-0.17	AA	-0.22	0.05	0.1155	0.7340
31.	-0.17	other	-0.49	0.32	0.3061	0.5801
30.	-1.12	EA	-0.71	-0.41	0.6443	0.4222
30.	-1.12	AA	-0.88	-0.24	0.1573	0.6917
30.	-1.12	other	-0.49	**-0.63**	0.7980	0.3717
34.	0.91	EA	0.65	0.26	0.3642	0.5462
34.	0.91	AA	0.48	0.43	0.7657	0.3816
34.	0.91	other	0.48	0.42	0.3828	0.5361
32.	0.33	EA	0.39	-0.07	0.4161	0.5189
32.	0.33	AA	0.61	-0.29	0.1486	0.6999
32.	0.33	other	0.49	-0.16	0.0086	0.9262

Note: *|DIF contrast| ≥ 0.5 logits in bold.* EA = European American, AA= African American.
**p-value* < .004

Validity. In order to assess the convergent validity of the ECBI scales, the extent to which the scores of the ODBTA, CPB, and IB five-point scales were correlated with measures of related constructs was examined. Specifically, it was expected that the ODBTA scale would positively correlate with the Oppositional Defiant Disorder Symptom Scale of the Conners Parent Rating Scale, 3rd Edition (CPRS). The CPB scale was expected to correlate positively with the Conduct Disorder Symptom Scale of the CPRS. Finally, the IB scale was expected to correlate positively with the ADHD Predominately Inattentive Symptom Scale of the CPRS. Furthermore, for the CPRS Content Scales, it was expected that the IB scale would positively correlate with the Inattention Content Scale of the CPRS, given the similar content of the two scales.

In order to assess the discriminant validity of the ECBI scales, the extent to which the scores of the ODBTA, CPB, and IB five-point scale were correlated with theoretically unrelated variables was examined. It was expected that the ODBTA, CPB, and IB scale

scores would not be strongly correlated with either the Peer Relations or the Learning Problems Content Scale scores of the CPRS.

Pearson's product moment correlation coefficients, *r,* were used to examine the relationships between the ODBTA, CPB, and IB scale scores and the CPRS scale scores. Pearson's *r* is a parametric test used to measure the strength of association between two variables and is appropriate for continuous variables. An *r* value of one indicates a perfect positive correlation, a value of negative one indicates a perfect negative correlation, and a value of zero indicates no correlation. In order to assess the strength of the correlation, the guidelines suggested by Evans (1996) were used (Table 29). Scale totals based on the five-point rating scale were used as the CFA results for the seven-point and the five-point rating scale structures were similar.

Table 29

Guidelines to Describe the Strength of Pearson's Correlation Coefficient (Evans, 1996)

Pearson's Correlation Coefficient Range	Strength
0.00 to 0.19	Very Weak
0.20 to 0.39	Weak
0.40 to 0.59	Moderate
0.60 to 0.79	Strong
0.80 to 1.00	Very Strong

Note: Adapted from "Straightforward statistics for the behavioral sciences*"* by J.D. Evans, 1996, Pacific Grove, CA: Brooks/Cole Publishing.

The correlation analyses included 194 of the cases with CPRS data out of the total sample (N = 221). The CPRS is appropriate for parents of children ages six to 18. Of the 27 excluded cases, 21 were excluded due age, i.e., younger than six years of age. The remaining six cases were excluded due to missing CPRS data. The subsample (n = 194) was 72.2% male and 27.8% female, with an average age of 9.8 years (range six to 17, *SD* = 2.67). A total of 67.5% of the parents in this subsample were married;

17.0% were divorced; 2.6% were separated; 9.3% were single and had never been married; 2.1% were living with someone; 1% were widowed; and one respondent's marital status was missing. Regarding ethnicity, 42.3% of the sample was Hispanic, 41.2% was European American, 12.9% was African American, and 3.6% identified as "other." The Pearson correlation coefficients for the relationships between the ODBTA, CPB, and IB five-point scale scores and the CPRS scale scores are shown in Table 30. The Pearson correlation coefficients for the relationships between the ODBTA, CPB, and IB seven-point scale scores and the CPRS scale scores are shown in table B1 of Appendix B.

Table 30

Pearson's Product Moment Correlation Coefficients for the Five-point Rating Scale

	Pearson's *r*		
	IB	ODBTA	CPB
CPRS Symptom Scales			
ADHD Predominately Inattentive Type	0.534**	0.331**	0.182**
ADHD Predominately Hyperactive-Impulsive Type	0.401**	0.502**	0.415**
Conduct Disorder	0.101	0.556**	0.561**
Oppositional Defiant Disorder	0.303**	0.676**	0.524**
CPRS Content Scales			
Peer Relations	0.158*	0.364**	0.276**
Aggression	0.151*	0.480**	0.485**
Learning Problems	0.217**	0.308**	0.294**
Executive Functioning	0.555**	0.294**	0.142*
Inattention	0.610**	0.291**	0.190**
Hyperactivity/Impulsivity	0.409**	0.487**	0.364**

Note: CPRS= Conners Parent Rating Scale, 3rd Edition*;* ECBI= Eyberg Child Behavior Inventory; IB = Inattentive Behaviors; ODBTA = Oppositional Defiant Behavior Toward Adults; and CPB = Conduct Problem Behavior.
**p-value <.05, ** p-value <.01*

The results in Table 30 show that there was a moderate, positive correlation between the IB scale score and the ADHD Predominately Inattentive Type Symptom scale score, i.e., $r = 0.534$, n = 194, $p < .01$, and a strong, positive correlation between the

IB scale score and the Inattention Content Scale score, i.e., $r = 0.610$, n = 194, $p < .01$. Additionally, there was a strong, positive correlation between the ODBTA scale score and the Oppositional Defiant Disorder Symptom scale score, i.e., $r = 0.676$, n = 194, $p < .01$. Finally, there was a moderate, positive correlation between the CPB scale score and the Conduct Disorder Symptom scale score, i.e., $r = 0.561$, n = 194, $p < .01$. These results provide evidence for the convergent validity of the three ECBI scales.

Regarding the discriminant validity of the IB, ODBTA, and CPB scales, the IB ($r = 0.158$, n = 194, $p < .05$); ODBTA ($r = 0.364$, n = 194, $p < .01$); and CPB ($r = 0.276$, n = 194, $p < .01$) scale scores were weakly correlated with the Peer Relations Content Scale score of the CPRS. Similarly, the IB ($r = 0.217$, n = 194, $p < .01$); ODBTA ($r = 0.308$, n = 194, $p < .01$); and CPB ($r = 0.294$, n = 194, $p < .01$) scale scores were weakly correlated with the Learning Problems Content Scale score of the CPRS. These results provide evidence for the discriminant validity of the three ECBI scales.

Hypothesis Three

The third hypothesis states that the ECBI would demonstrate adequate reliability within an ethnically diverse sample. In order to assess the reliability of the ODBTA, CPB, and IB scales, Cronbach's alpha and Rasch-based estimates of reliability, i.e., separation coefficients and reliability indices for both persons and items, were used (Bond & Fox, 2015; Linacre, 2017).

The separation coefficient is a "signal-to-noise" ratio where signal is the true variance and noise is the error variance (Linacre, 2017). Separation values less than two for persons suggest that the instrument may not be sensitive enough to differentiate between high and low responders. Low person separation may indicate that more items

are needed or that the person sample has too narrow of an ability range for meaningful measurement (Linacre, 2017). Item separation coefficients less than three may indicate that the person sample is not large enough to confirm the item difficulty hierarchy.

As shown in Table 31, the person separation coefficients for the ODBTA and IB scales were above the suggested guideline of two. However, the person separation coefficient for the CPB scale was 1.53. This is just below the suggested cutoff of two and indicates that the scale may not be sensitive enough to distinguish adequately between high and low performers. Adding more items to the CPB scale or more persons with varied ability ranges may improve the person separation.

The item separation coefficients for the ODBTA, CPB, and IB scales were above the suggested cutoff of three (Table 31). This implies that the person sample was large enough to confirm the item difficulty hierarchies for all three scales.

Table 31

Separation Coefficients and Reliability Indices

	Separation Coefficient		Reliability Index		Cronbach's Alpha
ECBI Scale	Person	Item	Person	Item	
ODBTA	3.03	3.27	0.91	0.91	0.92
CPB	1.53	7.47	0.70	0.98	0.78
IB	2.48	5.22	0.86	0.96	0.88

Note: Reliability statistics are based on both extreme and non-extreme measures.

The reliability indices are a reflection of the extent of reproducibility of the order of person and item measures. Reliability is the true variance divided by the observed variance, where observed variance = true variance + error variance. Reliability indices have a range of zero to one, and values less than 0.5 imply high measurement error. The Person Reliability index is analogous to Cronbach's alpha. Reliability indices and Cronbach's alpha values equal to or greater than 0.70 were deemed acceptable (Linacre,

2017). As Cronbach's alpha conventionally includes both extreme and non-extreme scores, the Rasch-based reliability indices reported also included both extreme and non-extreme measures.

The person and item reliability indices for the ODBTA, CPB, and IB scales as shown in Table 31 were above the suggested cutoff of 0.70 suggesting that the range of ability within the sample was adequate, that the item difficulty range was adequate, and that the sample was large enough for reproducibility of person and item measures (Furr, 2018; Linacre, 2017). Furthermore, Cronbach's alpha for the ODTBA, CPB, and IB scales were also above the suggested cutoff of 0.70.

Chapter 5: Discussion

The overall aim of this study was to assess the psychometric functioning of the ECBI in order to understand better the optimal interpretation of ECBI scores within a culturally diverse sample. Specifically, three hypotheses related to the psychometric properties of the ECBI were tested. First, it was hypothesized that the three-factor model (Burns and Patterson, 1991, 2000) of the ECBI would provide a better fit to the data than the one-factor model as demonstrated in other research (e.g., Axberg et al., 2008; Burns and Patterson, 1991, 2000; and Weis et al., 2005). Second, it was hypothesized that several of the intensity scale items, such as items related to externalizing behaviors, would function differently between Hispanic and non-Hispanic samples. Third, it was hypothesized that the ECBI scores would demonstrate adequate reliability within an ethnically diverse sample.

Hypotheses

The results of the CFA revealed that the three-factor model proposed by Burns and Patterson (1999, 2000) provided a better fit to the data compared to the one-factor model. These results are not unexpected given theoretical understandings of externalizing disorders which support distinctions between ADHD, ODD, and CD (Connor & Doerfler, 2008; Hinshaw, 1987). However, they do provide novel empirical support for the generalizability of the findings related to the factor structure of the ECBI to populations with a large Hispanic representation. Additionally, these findings support using the ECBI as a multi-dimensional measure which, as suggested by Burns and Patterson (2000), would increase its clinical and research utility.

Beyond the CFA findings that support the superiority of the three-factor model, there were noteworthy similarities between the results of Burns and Patterson's (2000) CFA of the three-factor model and the results of this study. Specifically, modification indices for both studies suggested that model fit would improve by correlating the error terms of items "verbally fights with sisters and brothers" with "physically fights with sisters and brothers," and "steals" with "lies." Although correlating error terms often improves model fit, there are important considerations to explore to ensure that modifications to the model are not uniquely fit-driven. For example, Brown (2015) cautions against the use of correlated error terms solely in an effort to improve model fit, as some modification indices can appear illogical, be the result of chance occurrences in the data or sample specific characteristics, and/or be indicative of an additional factor. Therefore, any modifications to the model, such as correlating error terms, should be justified by prior evidence or theory.

As rationale for correlating the error terms of item 25 (i.e., verbally fights with sisters and brothers) with item 27 (i.e., verbally fights with sisters and brothers) and of item 22 (i.e*.,* lies) with item 21 (i.e., steals) of the CPB scale, Burns and Patterson (2000) referred to the similar content of the dyads and the high co-occurrence of items 22 and 21. The justifications for model modifications specified by Burns and Patterson are also appropriate justifications for this study. In fact, similar evidence for model modification for this study was offered in the CFA results section of this document. Furthermore, the concerns when correlating error terms presented by Brown (2015) were considered. However, they were not applicable to this study since the modifications were not illogical, did not appear to be the result of chance occurrences or sample specific

characteristics, and were not indicative of an additional factor. Therefore, rather than creating a separate meaningful factor, correlating the errors of the three item pairs was appropriate.

Interestingly, in addition to the support presented, as well as the precedent set by Burns and Patterson's (2000) modification of the three-factor model, the results of the PCA of residuals provide additional justification for correlating the error terms of item 25 with 27 and item 13 (i.e., has temper tantrums) with 17 (i.e., yells or screams). Specifically, review of the contrast plot of residual loadings for the ODBTA scale (Figure 3) revealed that the pattern of residuals for items 13 and 17 clustered together, along with item 12, and contrasted with the pattern of residuals for other items in the scale. Similarly, for the CPB scale, the pattern of residual loadings (Figure 7) for items 25 and 27 clustered together and contrasted with the pattern of residuals for the other items in the scale, including items 22 and 21. Although contrasting clusters of residual loadings can sometimes be indicative of a meaningful secondary dimension, these data indicated that they are likely the result of shared content within the item dyads not found in the other items.

It is important to note that CFA and PCA of residuals are not to be interpreted the same (Linacre, 2017). However, these results tell a similar story related to the residual variance of the CPB and ODBTA items (Linacre, 2020). Both the CFA results and the PCA of residuals results suggest that although some items share residual variance, the covariance is reasonable, as the items share content within a larger dimension. Additionally, these findings highlight the potential benefits of using CFA and Rasch modeling to complement one another.

Rating scale functioning was assessed by evaluating the monotonicity of each scale. Specifically, the ordering of observed averages and Andrich thresholds was examined. Additionally, category mean-square statistics were used to estimate appropriate category usage. The rating scale functioning assessment revealed that a five-point rating scale optimized rating scale functioning for all three of the ECBI scales. This was not unexpected given that, for the seven-point rating scale, categories two and three are both labeled "seldom" and categories five and six are both labeled "often." Therefore, collapsing these categories improved the monotonicity of the scales.

The rating scale functioning analysis for the IB scale showed misfit for category one (infit MNSQ = 2.23). As previously mentioned, this is likely the result of ECBI scoring procedures which direct the examiner to select category one, i.e., "never," for missing responses (Linacre, 2017). Specifically, rules such as these can be problematic when evaluating rating scale functioning, as they introduce randomness and unexpected responses that can distort the rating scale. For example, for the IB scale, the reason for endorsing "never" can be different than the reason for not responding to an item. Ideally, in such cases where misfit is believed to be associated with scoring rules, re-evaluating the rating scale functioning with those unanswered items coded as "not administered," or omitting individual observations would allow for a better understanding of category usage. However, such an analysis was not possible, as the database coding structure did not differentiate which codes of "never" were a result of unanswered items.

Typically, a uniform distribution of observations across categories of a rating scale is ideal for step calibration. However, observation distributions are a reflection of the manifestation of a trait in a sample or population. Some traits, such as criminal

behaviors, are expected to have a skewed distribution (Linacre, 2002). Therefore, the trait that is measured by the rating scale is an important consideration when assessing rating scale functioning. For example, the rating scale functioning assessment of the CPB scale showed that category five, i.e., always, was infrequently used, while category one, i.e., never, was most frequently used. Additionally, given the severity of the behaviors associated with conduct problems, a right-tailed distribution is expected. So, for the CPB rating scale, the skewed distribution of observations is a reflection of the underlying trait and is not indicative of abnormal category usage.

Relating to item bias, while the items of the ODBTA, CPB, and IB scales did not exhibit statistically significant differences in item functioning, several items were found to have considerable values for DIF contrast. However, none of these considerable values resulted from contrasting the Hispanic group values with the European American group values. These were the largest two groups within the total sample, while the African American and the "other" groups had markedly smaller sample sizes, which presented a risk for committing type two errors associated with low statistical power. Given the notably small sample sizes of the "other" group and the African American group, the results related to item bias for those groups are presented as pilot data for consideration for future research (Linacre, 2017). In fact, even the largest two groups, the Hispanic and the European American groups, are considered small for DIF analyses with adequate power (Scott et al., 2009). Therefore, while these results provide preliminary evidence for the cross-cultural use of the ECBI, future studies using larger samples are needed to confirm further the item invariance for the ECBI scales and to generalize these findings.

In addition to the concerns associated with low statistical power, the clinical nature of the sample may have had implications related to response patterns that are important to consider when conceptualizing the DIF results. Specifically, Hispanic individuals who maintain collectivistic cultural values are less likely to rate externalizing symptoms as problematic (Schmitz & Velez, 2003). This cultural consideration provided support for the hypothesized item invariance between Hispanic and non-Hispanic groups. However, data for this study were obtained from a clinical database of individuals who perceived behaviors to be problematic and sought psychological services. Therefore, differences in item functioning associated with Hispanic collectivistic cultural values may be less likely to occur in clinical samples comprised of families who are already seeking psychological services in comparison to heterogenous samples comprised of clinical and non-clinical groups.

Finally, the three scales of the ECBI demonstrated acceptable reliability within a predominately Hispanic sample as indicated by reliability indices and Cronbach's alpha values >0.70. For the CPB scale, the person separation coefficient, i.e., 1.53, was below the expected value of two but still corresponded to a person "test" reliability value of 0.70. This suggests that the eight items of the CPB scale reliably discriminated between high and low performers. Since person separation is impacted by the range of person measures and the targeting of the person and item measures, increasing the range of the sample by adding more persons with varied abilities or adding more items can increase person separation. For the CPB scale, Figure 8 shows that adding more persons with higher trait levels would improve the person-item targeting and increase the range of the sample.

The reliability findings for the three ECBI scales speak to the reproducibility of the person and item measures and the internal consistency of the ODBTA, CPB, and IB item scores. As reliability is a necessary component in asserting the viability of the cross-cultural use of a measure, these results, in addition to those related to the item invariance noted above, support the use of the ECBI as a multi-dimensional measure in ethnically diverse populations (Van de Vijver & Poortinga, 2005).

Overall, these findings add to the extant research related to the psychometric properties of the ECBI and confirm the superiority of the three-factor model proposed by Burns and Patterson (2000). This study also provides novel support for the use of the three scales of the ECBI within Hispanic populations. Noteworthy are the results related to the optimization of rating scale functioning of the ECBI scales by using a five-category scale instead of a seven-category scale. To the author's knowledge, modifications to the ECBI's rating scale have not been proposed in prior research. Implications related to the clinical and research utility and cross-cultural use of the ECBI scales, as well as the limitations of this study and considerations for future research are discussed below.

A Multidimensional Measure

Use of the ECBI as a measure of three distinct domains of problematic behaviors in children and adolescents, rather than as a unidimensional measure of general problematic behaviors, can increase the assessment value and the utility of the ECBI across settings. First, the scales of the ECBI would allow providers in pediatric primary care and school settings to differentiate between oppositional defiant, inattentive, and problematic conduct behaviors, which would aid in early identification and more exact treatment referrals. For example, a child whose scores are higher for the IB scale

compared to the CPB and ODBTA scales, would likely benefit more from a referral for services to improve concentration and sustained attention, than from a referral for services with an emphasis on decreasing defiant or disruptive behaviors. Moreover, as previously mentioned, mental health disparities disproportionately impact minority populations, such as Hispanic individuals. The use of the ECBI scales can directly benefit these underserved populations, as Hispanic individuals are more likely to seek help for mental health concerns in settings such as primary care offices and as Hispanic children are more likely to be identified in school settings (Pagano et al., 2000).

Second, scale scores could be used to assess more accurately and to monitor more precisely behavior change throughout treatment. For example, ECBI total scores are currently used to monitor weekly progress in Parent-Child Interaction Therapy (PCIT; Brinkmeyer & Eyberg, 2003). However, as PCIT is an evidenced-based intervention for ODD, the ODBTA scale score would provide more meaningful indications of treatment progress than ECBI total scores. Utilizing the ECBI total scale score in such circumstances could potentially dilute or inaccurately augment indications of behavior change and, as a result, of treatment efficacy. Therefore, the use of the three ECBI scale scores could be helpful not only to assess behavior change better, but also to identify more easily target behaviors in treatment planning and intervention.

Third, ECBI scale scores can be especially useful in research procedures including screening activities as part of sample recruitment and group assignment. In conducting research, obtaining information from prospective participants is an early and important step in determining study eligibility. Depending on the focus of the study, the ECBI scale scores can be used to assess whether study inclusion or exclusion criteria are

met, as well as to assign participants to appropriate intervention groups. Additionally, being able to assess different domains of behavior can aid in the interpretation of research findings and in the consideration of potential confounding variables. Notably, the research enhancements provided by the three scales of the ECBI can extend beyond psychological research into pharmaceutical research related to medication interventions for ADHD symptoms.

Cross-Cultural Use

Beyond confirming the factor structure of the ECBI, these findings provide a better understanding of the ECBI's cross-cultural use. The Standards for Educational and Psychological Testing (2014) recommends a thorough psychometric evaluation in order to identify potential construct biases that may exist as result of cross-cultural differences (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). This study adheres to those recommendations and provides support for the structural and measurement equivalence of the ECBI in an ethnically diverse sample, which has not been explored in other research due to lack of culturally heterogenous samples.

It was hypothesized that some of the intensity scale items of the ECBI, specifically those relating to externalizing behaviors, would function differently between Hispanic and non-Hispanic samples. Culturally specific expectations for child behavior typically observed in Hispanic families were the bases for this hypothesis. Specifically, collectivistic cultural values that may facilitate a more accepting and understanding view of child behavior or that may make Hispanic parents less likely to rate externalizing behaviors as problematic contributed to this hypothesis (Borrego et al., 2006; Canino &

Guarnaccia, 1997; Halguenseth, Ispa, & Rudy, 2006; Schmitz & Velez, 2003). Nevertheless, none of the items across the ODTBA, IB, or CPB scales demonstrated significant DIF or had considerable values for DIF contrast when item functioning for the Hispanic group was compared to the European American group. However, several of the items demonstrated considerable values for DIF contrast without statistical significance when item functioning comparisons were made between the Hispanic and African American groups and the Hispanic and "other" groups.

Despite these findings, review of the items with considerable DIF did not indicate an overtly discernible pattern that would suggest that Hispanic cultural values influenced item functioning in any significant way(s). Additionally, while several of the items related to externalizing behavior problems, e.g., argues with adults, has temper tantrums, and physically fights with sisters and brothers, did demonstrate DIF contrast values above the expected threshold, $|\text{DIF contrast}| \geq 0.50$ logits, the notably small sample sizes for the African American and "other" groups, as well as the smaller than typical sample sizes for the Hispanic and European American groups limited the ability to draw definitive conclusions from these results. Therefore, the lack of clinically significant or considerable DIF contrast values for the comparisons of item functioning between the Hispanic and the European American groups provides preliminary support for the item invariance of the ECBI scales across these cultural groups.

The cross-cultural use of the ECBI can be supported further by additional explorations of its factor structure that include the 14 omitted items. Confirmation of the three-factor model proposed by Burns and Patterson (1991, 2000) was the basis for this study. The 22 items that the authors found to load on three meaningful factors were

included, while the 14 items with low factor loadings were dropped. However, given that a main criticism of the study by Burns and Patterson is the use of culturally homogenous samples, it could be argued that the factor structure of the ECBI using the original 36 items requires further exploration within culturally diverse samples. Therefore, future studies that utilize similar approaches to Burns and Patterson's exploratory factor analysis of the 36-item ECBI are needed using large, culturally diverse samples. Specifically, inclusion of the omitted items would be a worthy area of research considering that Hispanic family values, e.g., *respeto,* may impact the factor loadings of omitted items such as *hits parents, refuses to eat food presented,* and *interrupts.* Further exploratory analyses of the ECBI's factor structure in ethnically diverse samples may reveal that the omitted items could be retained in the three-factor model to improve the measurement of problematic behaviors in ethnically diverse populations.

Limitations of the Study

Sample characteristics such as sample size and the sample demographics, i.e., size of ethnic groups, as well as the clinical nature of the sample, were limitations of this study. As a result, comparisons of CFA results across ethnic groups, such as Hispanic, European American, and African American groups, were not feasible due to limited sample size. Such comparisons could have provided further support for the superiority of the three-factor structure of the ECBI across groups and added to the determination of the cross-cultural utility of the ECBI. Further, given the small size of the Hispanic, European American, African American, and "other" groups, results related to DIF were exploratory, rather than decisive, due to the potential for type two errors associated with low statistical power. While *a priori* statistical power analyses in order to estimate the

number of observations needed to improve the chances of detecting a true effect would have been ideal, the archival nature of the data used for this study limited such evaluations. Finally, the data for this study were obtained from a clinical sample of families who were presenting for psychological services, which may limit the generalizability of the findings to non-clinical populations in which the ECBI scales may be used for screening purposes. Additionally, the possible implications of the use of a clinical sample on differential item functioning analyses, which have been discussed, are further limitations in the determination of item invariance across ethnic groups.

There were also methodological limitations for this study, such as the data that were available in the archival database and how ethnicity was recorded in the database. In cross-cultural research, acculturation is an important factor to consider when exploring the relationship between culture and a variety of constructs such as parenting practices, response patterns, and health behaviors (Fox, Thayer, & Wadhwa, 2017). However, the database used in this study did not have acculturation assessment data. Additionally, ethnic categories were limited to "Hispanic" for individuals of Latin or Hispanic descent, and information related to country of origin was not available. Furthermore, as previously mentioned, the database coding structure did not indicate which codes of "never" were the result of unanswered items which restricted the analysis of the effect(s) of scoring directions on rating scale functioning.

Implications for Future Research

Future studies are needed to replicate the findings of this study, to improve upon the limitations described above, and to establish norming criteria. Replication is needed in order to provide evidence for the generalizability of these results beyond this sample.

Additionally, as previously mentioned, exploratory studies to evaluate the structural and measurement equivalence of the 36-item ECBI in culturally diverse samples are warranted. Relating to reliability, the ODBTA and IB scales demonstrated good item and person separation; however, the person separation for the CPB scale was just below the suggested guideline. Replicating the results of this study in a sample with a wider range of person measures, would aid in providing further clarity regarding the ability of the CPB to differentiate between high and low child ratings. Additionally, further investigation of the rating scale functioning is needed in order to make decisive conclusions about the optimal number of categories for the ECBI scales. Furthermore, although not the aim of this study, the comprehensive analytic steps taken prompted consideration of possible modifications to the ECBI that would increase measurement precision beyond those discussed above. For example, adding a "not applicable" or "no response" option could be piloted in future research in order to clarify the meaning of category one and to improve rating scale functioning.

Relating to the aforementioned limitations, several recommendations for future studies related to the sample and to the methodology, e.g., the data collected for analysis, are warranted. First, future investigations of the cross-cultural equivalence of the ECBI scales should seek larger sample sizes with robust subsamples across all ethnic groups and for cultural differentiation within ethnic subgroups. Allowing for differentiation of ethnicity based on nation of origin is recommended. Although a common practice, utilizing terms such as "Hispanic" increases the risk of overlooking important cultural distinctions that may be meaningful to cross-cultural research. Second, samples that include both clinical and non-clinical groups are recommended in order to understand

better the functioning of the ECBI scale items across ethnic groups. Third, statistical power analyses prior to participant recruitment and data collection are recommended to help identify necessary sample sizes for detecting specific effect sizes and decreasing the chance of type two errors in future studies. Fourth, incorporating an acculturation measure during data collection is recommended to understand better the needs of minority groups and to account more fully for possible confounding effects of culture. Finally, consideration of caregiver factors such as stress and/or psychopathology, as well as incorporation of paternal report, if applicable, is suggested to explore the potential impact, if any, of these factors on rating scale functioning.

Finally, in order to move toward the implementation and utilization of the ECBI scales, future studies which include norming procedures are needed. Norms for each proposed scale of the ECBI using a five-point rating paradigm are needed in order to make inferences about a child's scale score compared to that of others. Consideration of the need for sample-specific norms, or norms associated with acculturation levels, should be explored. However, should future studies also support the structural and measurement equivalence of the ECBI scales, the need for sample-specific norms may be obviated.

Conclusion

As the population of the United States continues to diversify, cross-cultural measurement equivalence becomes a more salient issue in assessment. This study illustrates the thorough psychometric evaluation that is needed in order to establish the appropriate cross-cultural use of measurement tools. Historically, the assessment of measurement equivalence has most commonly involved classical test theory approaches, including exploratory and confirmatory factor analyses (Van de Vijver & Poortinga,

2005). However, modern test theory approaches, such as Rasch modeling, have grown in popularity (Byrne et al., 2009). In this study, the benefits of complementing more traditionally used approaches, such as CFA, with Rasch modeling was highlighted. Specifically, Rasch modeling allows for reliability and item functioning assessments beyond the sample-dependent information provided by CFA. Alternatively, factor analytic approaches are often more familiar to researchers than Rasch modeling and may be more practical given the sample size requirements of Rasch modeling. Therefore, utilizing techniques from both methodologies, as was demonstrated in this study, may be the best approach for comprehensive analytic strategies (Cappelleri et al., 2015; Kean & Reilly, 2004).

In conclusion, while further evaluation of the proposed three-factor ECBI is necessary prior to large scale implementation, professionals are encouraged to consider the possible improvements to the utility of the ECBI afforded by using scale scores versus a total score. Depending on the intended use, evaluators may find that using the ECBI as a measure of three distinct dimensions of problematic behaviors is not only well-supported by the research literature (e.g., Axberg et al., 2008; Burns & Patterson, 1999, 2000; Stern, 2007; Weis, Lovejoy, & Lundahl, 2005) but also can result in more meaningful assessment data than use of a total score.

References

Abdi, H. (2007). The Bonferroni and Sidak corrections for multiple comparisons. In N. Salkin (Ed.), *Encyclopedia of measurement and statistics*. Thousand Oaks, CA: Sage Publications.

Abidin, R. R. (1992). The determinants of parenting behavior. *Journal of Clinical Child Psychology*, *21*(4), 407-412. doi: 10.1207/s15374424jccp2104_12

Abidin, R. R. (2012). *Manual for the Parenting Stress Index* (4th ed). Odessa, FL: Psychological Assessment Resources.

Abraído-Lanza, A. F., Echeverría, S. E., & Flórez, K. R. (2016). Latino immigrants, acculturation, and health: Promising new directions in research. *Annual Review of Public Health*, *37*, 219-236. doi: 10.1146/annurev-publhealth-032315-021545

Acevedo-Polakovich, I. D., Crider, E. A., Kassab, V. A., & Gerhart, J. I. (2011). Increasing service parity through organizational cultural competence. In L.P. Buki & L.M. Piedra (Eds.), *Creating infrastructures for Latino mental health* (pp. 79-98). doi: 10.1007/978-1-4419-9452-3_1

Achenbach, T. M. (2001). What are norms and why do we need valid ones? *Clinical Psychology: Science and Practice,* 8(4), 446-450. doi: doi.org/10.1093/clipsy.8.4.446

Achenbach, T. M. (2009). *The Achenbach system of empirically based assessment (ASEBA): Development, findings, theory, and applications*. Burlington, VT: University of Vermont Research Center for Children, Youth, & Families.

Achenbach, T. M. (2017). Future directions for clinical research, services, and training: Evidence-based assessment across informants, cultures, and dimensional

hierarchies. *Journal of Clinical Child and Adolescent Psychology*, *46*(1), 159-169. doi: 10.1080/15374416.2016.1220315

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*(3), 251-275. doi: 10.1111/j.1469-7610.2007.01867.x

Achenbach, T. M., & Rescorla, L. A. (2000). *Manual for the ASEBA preschool forms & profiles*. Burlington: University of Vermont Research Center for Children, Youth, and Families.

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA school-age forms & profiles*. Burlington: University of Vermont Research Center for Children, Youth, and Families

Achenbach, T. M., & Rescorla, L. A. (2015). *Multicultural supplement to the manual for the ASEBA adult forms & profiles*. Burlington: University of Vermont Research Center for Children, Youth, and Families.

Achenbach, T. M., & Ruffle, T. M. (2000). The Child Behavior Checklist and related forms for assessing behavioral/emotional problems and competencies. *Pediatrics in Review*, *21*(8), 265-271. doi: 10.1177/1063426611434158

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*(2), 213. doi: 10.1037/0033-2909.101.2.213

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*(1), 67-91. Retrieved from http://www.jstor.org.ezproxylocal.library.nova.edu/stable/1434777

Alegria, M., Vallas, M., & Pumariega, A. J. (2010). Racial and ethnic disparities in pediatric mental health. *Child and Adolescent Psychiatric Clinics of North America*, *19*(4), 759-774. doi: 10.1016/j.chc.2010.07.001

Allison, C., Baron-Cohen, S., Stone, M. H., & Muncer, S. J. (2015). Rasch modeling and confirmatory factor analysis of the systemizing quotient-revised (SQ-R) scale. *The Spanish Journal of Psychology*, *18*. doi: 10.1017/sjp.2015.19

Allison, C., Baron-Cohen, S., Wheelwright, S. J., Stone, M. H., & Muncer, S. J. (2011). Psychometric analysis of the Empathy Quotient (EQ). *Personality and Individual Differences, 51*, 829-835. doi:10.1016/j.paid.2011.07.005

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. (1987). *Diagnostic and statistical manual of mental disorders: 3rd edition, text rev.* Washington, DC: American Psychiatric Association.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5*. Washington, D.C: American Psychiatric Association.

American Psychological Association (2014). The Standards for Educational and Psychological Testing. Retrieved from http://www.apa.org/science/programs/testing/standards.aspx

Andrich, D. (1978). A rating scale formulation for ordered response categories. *Psychometrika, 43,* 561-573. doi: 10.1007/BF02293814

Andrich, D. (2006). Item discrimination and Rasch-Andrich thresholds revisited. *Rasch Measurement Transactions*, *20*(6), 1055-1057. Retrieved from https://www.rasch.org/rmt/rmt202a.htm

Arbuckle, J. L. (2017). Amos (Version 25.0) [Computer Program]. Chicago: IBM SPSS.

Axberg, U. L. F., Johansson Hanse, J. A. N., & Broberg, A. G. (2008). Parents' description of conduct problems in their children–A test of the Eyberg Child Behavior Inventory (ECBI) in a Swedish sample aged 3–10. *Scandinavian Journal of Psychology*, *49*(6), 497-505. doi: 10.1111/j.1467-9450.2012.00955.x

Bachman, J. G., O'Malley, P. M., & Freedman-Doan, P. (2010). Response styles revisited: Racial/ethnic and gender differences in extreme responding (Monitoring the Future Occasional Paper No. 72). Ann Arbor, MI: Institute for Social Research. Retrieved from: http://www.monitoringthefuture.org/

Barkley, R. A. (1997). Behavioral inhibition, sustained attention, and executive functions: Constructing a unifying theory of ADHD. *Psychological Bulletin*, *121*(1), 65. doi: 10.1037/0033-2909.121.1.65

Batchelor, J. H., & Miao, C. (2016). Extreme response-style: A meta-analysis. *Journal of Organizational Psychology*, *16*(2), 51. Retrieved from

https://www.researchgate.net/publication/316820164_Extreme_Response_Style_A_Meta-Analysis

Beal, A. C. (2004). Policies to reduce racial and ethnic disparities in child health and health care. *Health Affairs*, *23*(5), 171-179. doi: 10.1377/hlthaff.23.5.171

Belvedere, S. L., & de Morton, N. A. (2010). Application of Rasch analysis in health care is increasing and is applied for variable reasons in mobility instruments. *Journal of Clinical Epidemiology*, *63*(12), 1287-1297. doi: 10.1016/j.jclinepi.2010.02.012

Berkovits, M. D., O'Brien, K. A., Carter, C. G., & Eyberg, S. M. (2010). Early identification and intervention for behavior problems in primary care: A comparison of two abbreviated versions of parent-child interaction therapy. *Behavior Therapy*, *41*(3), 375-387. doi: 10.1016/j.beth.2009.11.002

Bernal, G., & Sáez‐Santiago, E. (2006). Culturally centered psychosocial interventions. *Journal of Community Psychology*, *34*(2), 121-132. doi: 10.1002/jcop.20096

Bezdjian, S., Krueger, R. F., Derringer, J., Malone, S., McGue, M., & Iacono, W. G. (2011). The structure of DSM-IV ADHD, ODD, and CD criteria in adolescent boys: A hierarchical approach. *Psychiatry Research*, *188*(3), 411-421. doi: 10.1016/j.psychres.2011.02.027

Bingham, C. R., Loukas, A., Fitzgerald, H. E., & Zucker, R. A. (2003). Parental ratings of son's behavior problems in high-risk families: Convergent validity, internal structure, and interparent agreement. *Journal of Personality Assessment*, *80*(3), 237-251. doi: 10.1207/S15327752JPA8003_03

Bond, T., & Fox, C. M. (2015). Model fit and unidimensionality. In T. G. Bond, & C. M. Fox (Eds.), *Applying the Rasch model: Fundamental measurement in the human sciences, third edition*. New York: Routledge.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, *8*, 3-62.

Boggs, S. R., Eyberg, S., & Reynolds, L. A. (1990). Concurrent validity of the Eyberg Child Behavior Inventory. *Journal of Clinical Child Psychology*, *19*(1), 75-78. doi: 10.1207/s15374424jccp1901_9

Borrego, J., Anhalt, K., Terao, S. Y., Vargas, E. C., & Urquiza, A. J. (2006). Parent-child interaction therapy with a Spanish-speaking family. *Cognitive and Behavioral Practice*, *13*(2), 121-133. doi: 10.1016/j.cbpra.2005.09.001

Brestan, E. V., Jacobs, J. R., Rayfield, A. D., & Eyberg, S. M. (2000). A consumer satisfaction measure for parent-child treatments and its relation to measures of child behavior change. *Behavior Therapy*, *30*(1), 17-30. doi: 10.1016/S0005-7894(99)80043-4

Bridges, A. J., Andrews III, A. R., Villalobos, B. T., Pastrana, F. A., Cavell, T. A., & Gomez, D. (2014). Does integrated behavioral health care reduce mental health disparities for Latinos? Initial findings. *Journal of Latina/o Psychology*, *2*(1), 37. doi: 10.1037/lat0000009

Brinkmeyer, M. Y., & Eyberg, S. M. (2003). Parent-child interaction therapy for oppositional children. In A. E. Kazdin & J. R. Weisz (Eds.), *Evidence-based psychotherapies for children and adolescents* (pp. 204-223). New York: Guilford.

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

Bruder, M. B. (2000). Family-centered early intervention: Clarifying our values for the new millennium. *Topics in Early Childhood Special Education*, *20*(2), 105-115. doi: 10.1177/027112140002000206

Burke, J. D., Hipwell, A. E., & Loeber, R. (2010). Dimensions of oppositional defiant disorder as predictors of depression and conduct disorder in preadolescent girls. *Journal of the American Academy of Child & Adolescent Psychiatry*, 49, 484-492. doi: 10.1016/j.jaac.2010.01.016

Burke, J. D., Loeber, R., & Birmaher, B. (2002). Oppositional defiant disorder and conduct disorder: A review of the past 10 years, part II. *Journal of the American Academy of Child & Adolescent Psychiatry*, *41*(11), 1275-1293.doi: 10.1097/01.CHI.0000024839.60748.E8

Burns, G. L., & Patterson, D. R. (1990). Conduct problem behaviors in a stratified random sample of children and adolescents: New standardization data on the Eyberg Child Behavior Inventory. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 2*(4), 391-397.doi: 10.1037/1040-3590.2.4.391

Burns, G. L., & Patterson, D. R. (1991). Factor structure of the Eyberg Child Behavior Inventory: Unidimensional or multidimensional measure of disruptive behavior? *Journal of Clinical Child and Adolescent Psychology*, *20*(4), 439-444. doi: 10.1207/s15374424jccp2004_13

Burns, G. L., & Patterson, D. R. (2000). Factor structure of the Eyberg Child Behavior Inventory: A parent rating scale of oppositional defiant behavior toward adults, inattentive behavior, and conduct problem behavior. *Journal of Clinical Child Psychology, 29*(4), 569-577. doi: 10.1207/S15374424JCCP2904_9

Burns, G. L., & Patterson, D. R. (2001). Normative data on the Eyberg Child Behavior Inventory and Sutter-Eyberg Student Behavior Inventory: Parent and teacher rating scales of disruptive behavior problems in children and adolescents. *Child & Family Behavior Therapy, 2*3(1), 15-28. doi: 10.1300/J019v23n01_02

Burns, G. L., Patterson, D. R., Nussbaum, B. R., & Parker, C. M. (1991). Disruptive behaviors in an outpatient pediatric population: Additional standardization data on the Eyberg Child Behavior Inventory. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *3*(2), 202-207. doi: 10.1037/1040-3590.3.2.202

Butler, A. M. (2011). Cross-racial measurement equivalence of the Eyberg Child Behavior Inventory factors among low-income young African American and Non-Latino White children. *Assessment*, *20*(4), 484-495. doi: 10.1177/1073191111431341

Byrne, B. M. (1994). *Structural equation modeling with EQS and EQS/Windows*. California: Sage Publications.

Byrne, B. M. (2005). Factor analytic models: Viewing the structure of an assessment instrument from three perspectives. *Journal of Personality Assessment*, *85*(1), 17-32. doi: 10.1207/s15327752jpa8501_02

Byrne, B. M., Oakland, T., Leong, F. T., van de Vijver, F. J., Hambleton, R. K., Cheung, F. M., & Bartram, D. (2009). A critical analysis of cross-cultural research and testing practices: Implications for improved education and training in psychology. *Training and Education in Professional Psychology*, *3*(2), 94. doi: 10.1037/a0014516

Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (2nd ed.). New York: Routledge. doi: 10.4324/9780203805534

Calzada, E. J., Fernandez, Y., & Cortes, D. E. (2010). Incorporating the cultural value of *respeto* into a framework of Latino parenting. *Cultural Diversity & Ethnic Minority Psychology*, *16*(1), 77–86. doi: 10.1037/a0016071

Canino, G., & Guarnaccia, P. (1997). Methodological challenges in the assessment of Hispanic children and adolescents. *Applied Developmental Science*, *1*(3), 124-134. doi: 10.1207/s1532480xads0103_3

Cano, M. Á., Schwartz, S. J., Castillo, L. G., Romero, A. J., Huang, S., Lorenzo-Blanco, E. I., ... & Lizzi, K. M. (2015). Depressive symptoms and externalizing behaviors among Hispanic immigrant adolescents: Examining longitudinal effects of cultural stress. *Journal of Adolescence*, *42*, 31-39. doi: 10.1016/j.adolescence.2015.03.017

Cappelleri, J. C., Lundy, J. J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for the quantitative assessment of items in developing patient-reported outcomes measures. *Clinical Therapeutics*, *36*(5), 648-662. doi: 10.1016/j.clinthera.2014.04.006

Cauffman, E., & MacIntosh, R. (2006). A Rasch differential item functioning analysis of the Massachusetts Youth Screening Instrument: Identifying race and gender differential item functioning among juvenile offenders. *Educational and Psychological Measurement*, *66*(3), 502-521. doi:10.1177/0013164405282460

Centers for Disease Control and Prevention (2013). *Mental Health Surveillance Among Children – United States, 2005-2011*. Retrieved from https://www.cdc.gov/mmwr/preview/mmwrhtml/su6202a1.htm?s_cid=su6202a1_w

Chamberlain, P., & Smith, D. K. (2003). Antisocial behavior in children and adolescents: The Oregon multidimensional treatment foster care model. In A. E. Kazdin & J. R. Weisz (Eds.), *Evidence-based psychotherapies for children and adolescents* (pp. 282-300). New York: Guilford.

Chafouleas, S. M., Kilgus, S. P., & Wallach, N. (2010). Ethical dilemmas in school-based behavioral screening. *Assessment for Effective Intervention, 35*(4), 245-252. doi: 10.1177/1534508410379002

Clarke, I. (2000). Extreme response style in cross-cultural research: An empirical investigation. *Journal of Social Behavior and Personality*, *15*(1), 137. Retrieved from https://search-proquest-com.ezproxylocal.library.nova.edu/docview/1292246471?pq-origsite=gscholar

Coffey, D. M., Javier, J. R., & Schrager, S. M. (2015). Preliminary validity of the Eyberg Child Behavior Inventory with Filipino immigrant parents. *Child & Family Behavior Therapy*, *37*(3), 208-223. doi: 10.1080/07317107.2015.1071978

Colvin, A., Eyberg, S., & Adams, C. (1999). Restandardization of the Eyberg Child Behavior Inventory. *Unpublished manuscript*. Gainesville, Florida: University of Florida, Child Study Laboratory.

Conners, C. K. (1970). Symptom patterns in hyperkinetic, neurotic, and normal children. *Child Development*, 667-682. doi: 10.2307/1127215

Conners, C. K. (2008). *Conners 3rd edition: Manual* (Vol. 14). Toronto, Ontario, Canada: Multi-Health Systems.

Conners, C. K., Sitarenios, G., Parker, J. D., & Epstein, J. N. (1998). The revised Conners' Parent Rating Scale (CPRS-R): Factor structure, reliability, and criterion validity. *Journal of Abnormal Child Psychology*, *26*(4), 257-268. doi: 10.1023/A:1022602400621

Connor, D. F., & Doerfler, L. A. (2008). ADHD with comorbid oppositional defiant disorder or conduct disorder: Discrete or nondistinct disruptive behavior disorders? *Journal of Attention Disorders*, *12*(2), 126-134. doi: 10.1177/1087054707308486d

Conway, J. M., & Huffcutt, A. I. (2003). A review and evaluation of exploratory factor analysis practices in organizational research. *Organizational Research Methods*, *6*(2), 147-168. doi:10.1177/1094428103251541

Costello, E. J., Mustillo, S., Erkanli, A., Keeler, G., & Angold, A. (2003). Prevalence and development of psychiatric disorders in childhood and adolescence. *Archives of General Psychiatry*, *60*(8), 837-844. doi: 10.1001/archpsyc.60.8.837

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical*

Assessment, Research & Evaluation, *10*(7), 1-9. Retrieved from http://www.pareonline.net/pdf/v10n7.pdf

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, *6*(4), 475-494. Retrieved from: http://journals.sagepub.com.ezproxylocal.library.nova.edu/doi/pdf/10.1177/00131 6444600600405

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334. Retrieved from http://hbanaszak.mjr.uw.edu.pl/TempTxt/Cronbach_ 1951_Coefficient%20alpha%20and%20the%20internal%20structure%20of%20te sts.pdf

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281. doi: 10.1037/h0040957

Davenport, E. C., Davison, M. L., Liou, P. Y., & Love, Q. U. (2015). Reliability, dimensionality, and internal consistency as defined by Cronbach: Distinct albeit related concepts. *Educational Measurement: Issues and Practice*, *34*(4), 4-9.

De Los Reyes, A. (2011). Introduction to the special section: More than measurement error: Discovering meaning behind informant discrepancies in clinical assessments of children and adolescents. *Journal of Clinical Child & Adolescent Psychology*, *40*(1), 1-9. doi: 10.1080/15374416.2011.533405

De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and

recommendations for further study. *Psychological Bulletin*, *131*(4), 483-509. doi: 10.1037/0033-2909.131.4.483

Dettlaff, A. J., & Johnson, M. A. (2011). Child maltreatment dynamics among immigrant and US born Latino children: Findings from the National Survey of Child and Adolescent Well-being (NSCAW). *Children and Youth Services Review*, *33*(6), 936-944. doi:10.1016/j.childyouth.2010.12.017

Dinh, K. T., Roosa, M. W., Tein, J. Y., & Lopez, V. A. (2002). The relationship between acculturation and problem behavior proneness in a Hispanic youth sample: A longitudinal mediation model. *Journal of Abnormal Child Psychology*, *30*(3), 295-309. doi: 0.1023/A:1015111014775

Dirks, M. A., De Los Reyes, A., Briggs‐Gowan, M., Cella, D., & Wakschlag, L. S. (2012). Annual research review: Embracing not erasing contextual variability in children's behavior–theory and utility in the selection and use of methods and informants in developmental psychopathology. *Journal of Child Psychology and Psychiatry*, *53*(5), 558-574. doi: 10.1111/j.1469-7610.2012.02537.x

Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education*, *37*(9), 830-837. doi: 10.1046/j.1365-2923.2003.01594.x

Duhig, A. M., Renk, K., Epstein, M. K., & Phares, V. (2000). Interparental agreement on internalizing, externalizing, and total behavior problems: A meta‐analysis. *Clinical Psychology: Science and Practice*, *7*(4), 435-453. doi: 10.1093/clipsy.7.4.435

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, *16*(1), 5-18. doi: 0.1007/s11136-007-9198-0

Embretson, S. E., & Reise, S. P. (2000). *Multivariate applications books series. Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Ennis, S. R., Rios-Vargas, M., & Albert, N. G. (2011). The Hispanic population: 2010 (2010 Briefs). Washington, DC: US Census Bureau.

Eklund, K., Renshaw, T. L., Dowdy, E., Jimerson, S. R., Hart, S. R., Jones, C. N., & Earhart, J. (2009). Early identification of behavioral and emotional problems in youth: Universal screening versus teacher-referral identification. *California School Psychologist*, *14*, 89-95. Retrieved from http://files.eric.ed.gov.ezproxylocal.library.nova. edu/fulltext/EJ878355.pdf

Escobar, J. I., Burnam, M. A., Karno, M., Forsythe, A., & Golding, J. M. (1987). Somatization in the community: Relationship to disability and use of services. *Archives of General Psychiatry*, *44*(8), 837-840. doi:10.1001/archpsyc.1987.01800200039006

Evans, J. D. (1996). *Straightforward statistics for the behavioral sciences*. Pacific Grove, CA: Brooks/Cole Publishing.

Evans, S. W., Owens, J. S., & Bunford, N. (2014). Evidence-based psychosocial treatments for children and adolescents with attention-deficit/hyperactivity disorder. *Journal of Clinical Child & Adolescent Psychology*, *43*(4), 527-551. doi: 10.1016/j.cpr.2006.01.002

Eyberg, S. M., Boggs, S. R., & Rodriguez, C. M. (1992). Relationships between maternal parenting stress and child disruptive behavior. *Child &Family Behavior Therapy*, *14*(4), 1-9. doi: 10.1300/J019v14n04_01

Eyberg, S. M., Funderburk, B. W., Hembree-Kigin, T. L., McNeil, C. B., Querido, J. G., & Hood, K. K. (2001). Parent-child interaction therapy with behavior problem children: One and two year maintenance of treatment effects in the family. *Child & Family Behavior Therapy*, *23*(4), 1-20. doi: 10.1300/J019v23n04_01

Eyberg, S. M., Nelson, M. M., & Boggs, S. R. (2008). Evidence-based psychosocial treatments for children and adolescents with disruptive behavior. *Journal of Clinical Child & Adolescent Psychology*, *37*(1), 215-237. doi: 10.1080/15374410701820117

Eyberg, S. M., & Pincus, D. (1999). *Eyberg Child Behavior Inventory & Sutter-Eyberg Student Behavior Inventory-Revised: Professional manual.* Odessa, FL. Psychological Assessment Resources.

Eyberg, S. M., & Ross, A. W. (1978). Assessment of child behavior problems: The validation of a new inventory. *Journal of Clinical Child & Adolescent Psychology*, *7*(2), 113-116. doi: 10.1080/15374417809532835

Eyberg, S. M., & Robinson, E. A. (1983). Conduct problem behavior: Standardization of a behavioral rating scale with adolescents. *Journal of Clinical Child & Adolescent Psychology*, *12*(3), 347-354. doi: 10.1080/15374418309533155

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272-299. doi: 10.1037/1082-989X.4.3.272

Fernández, D. P. R., Gorostiza, G. E., Lafuente, M. P., Ojembarrena, M. E., & Olaskoaga, A. A. (1998). Versión Española del ECBI (Eyberg Child Behavior Inventory): Medida de validez [Spanish version of ECBI (Eyberg Child Behavior Inventory): Measurement of validity]. *Atencion Primaria*, *21*(2), 65-74. Retrieved from http://www.elsevier.es/en-revista-atencion-primaria-27-resumen-version-espanola-del-ecbi-eyberg-15016#affa

Flores, G., Fuentes-Afflick, E., Barbot, O., Carter-Pokras, O., Claudio, L., Lara, M., ... & Valdez, R. B. (2002). The health of Latino children: Urgent priorities, unanswered questions, and a research agenda. *Journal of the American Medical Association*, *288*(1), 82-90. doi: 10.1001/jama.288.1.82

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, *7*(3), 286-299. doi: 10.1037/1040-3590.7.3.286

Fox, M., Thayer, Z., & Wadhwa, P. (2017). Assessment of acculturation. *Social Science and Medicine, 176,* 123-132. doi: 10.1016/j.socscimed.2017.01.029

Foy, J. M., Kelleher, K. J., Laraque, D., & American Academy of Pediatrics Task Force on Mental Health. (2010). Enhancing pediatric mental health care: Strategies for preparing a primary care practice. *Pediatrics*, *125*(Supplement 3), S87-S108. doi:10.1542/peds.2010-0788E

Funderburk, B. W., Eyberg, S. M., Rich, B. A., & Behar, L. (2003). Further psychometric evaluation of the Eyberg and Behar rating scales for parents and teachers of preschoolers. *Early Education and Development*, *14*(1), 67-82. doi: 10.1207/s15566935eed1401_5

Furr, R. M. (2018). *Psychometrics: An introduction* (3rd ed.). Thousand Oaks, CA: SAGE Publications, Inc.

Gall, G., Pagano, M. E., Desmond, M. S., Perrin, J. M., & Murphy, J. M. (2000). Utility of psychosocial screening at a school-based health center. *Journal of School Health*, *70*(7), 292-298.

Garcia-Tornel, S., Calzada, E. J., Eyberg, S.M., Alguacil, J.C., Serra, C.V., Mendoza, C.B.,…Domenech, A.T. (1998). Inventario Eyberg del comportamiento en niños: Normalización de la versión española y su utilidad para el pediatra extrahospitalario [Eyberg Child Behavior Inventory: Standardization of the Spanish version and its utility in pediatric practice]. *Anales Españoles de Pediatria*, *9*(48), 475-482. Retrieved from http://www.aeped.es/sites/default/files/anales/48-5-5.pdf

Garver, M. S., & Mentzer, J. T. (1999). Logistics research methods: Employing structural equation modeling to test for construct validity. *Journal of Business Logistics*, *20*(1), 33-57. Retrieved from https://search-proquest-com.ezproxylocal.library.nova.edu /docview/212605730/fulltext/47921FA4ACB6453CPQ/1?accountid=6579

Geisinger, K. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing normative interpretation of assessment instruments. *Psychological Assessment 6*(4), 304-312. doi: 10.1037/1040-3590.6.4.304

George, D., & Mallery, M. (2010). SPSS for Windows Step by Step: A Simple Guide and Reference, 17.0 update (10a ed.) Boston: Pearson.

Gerbing, D. W., & Anderson, J. C. (1988). An updated paradigm for scale development incorporating unidimensionality and its assessment. *Journal of Marketing Research*, 186-192. Retrieved from http://www.jstor.org.ezproxylocal.library.nova.edu/stable/3172650

Gessaroli, M.E., & de Champlain, A.F. (2005) Test Dimensionality: Assessment of. In B.S. Everitt, & D.C. Howell (Eds.), *Encyclopedia of statistics in behavioral science, volume four* (pp. 2014-2021). Hoboken, NJ: Wiley. doi: 10.1002/0470013192.bsa027

Glas, C. A. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika*, *53*(4), 525-546. Retrieved from http://hbanaszak.mjr.uw.edu.pl/TempTxt/Glas_1988_TheDerivationOfTestsForRaschMFromMultinomialDistribution.pdf

Goldman, L. S., Genel, M., Bezman, R. J., & Slanetz, P. J. (1998). Diagnosis and treatment of attention-deficit/hyperactivity disorder in children and adolescents. *Journal of the American Medical Association*, *279*(14), 1100-1107. doi: 0.1001/jama.279.14.1100

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry*, *38*(5), 581-586. doi: 10.1111/j.1469-7610.1997.tb01545.x

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*(1), 157-166. doi: 10.1037/a0018124

Grisso, T., Barnum, R., Fletcher, K. E., Cauffman, E., & Peuschold, D. (2001). Massachusetts Youth Screening Instrument for mental health needs of juvenile justice youths. *Journal of the American Academy of Child & Adolescent Psychiatry*, *40*(5), 541-548. doi: 10.1097/00004583-200105000-00013

Gross, D., Fogg, L., Webster-Stratton, C., Garvey, C., Julion, W., & Grady, J. (2003). Parent training of toddlers in day care in low-income urban communities. *Journal of Consulting and Clinical Psychology*, *71*(2), 261-278. doi: 10.1037/0022-006X.71.2.261

Gross, D., Fogg, L., Young, M., Ridge, A., Cowell, J., Sivan, A., & Richardson, R. (2007). Reliability and validity of the Eyberg Child Behavior Inventory with African–American and Latino parents of young children. *Research in Nursing & Health, 30*(2), 213-223. doi: /doi.org/10.1002/nur.20181

Guralnick, M. J. (2011). Why early intervention works: A systems perspective. *Infants and Young Children*, *24*(1), 6-28. doi: 10.1097/IYC.0b013e3182002cfe

Haack, L. M., & Gerdes, A. C. (2011). Functional impairment in Latino children with ADHD: Implications for culturally appropriate conceptualization and measurement. *Clinical Child and Family Psychology Review*, *14*(3), 318-328. doi: 10.1007/s10567-011-0098-z

Haack, L. M., Gerdes, A. C., Schneider, B. W., & Hurtado, G. D. (2011). Advancing our knowledge of ADHD in Latino children: Psychometric and cultural properties of

Spanish-versions of parental/family functioning measures. *Journal of Abnormal Child Psychology*, *39*(1), 33-43. doi: 10.1007/s10802-010-9441-y

Halgunseth, L. C., Ispa, J. M., & Rudy, D. (2006). Parental control in Latino families: An integrated review of the literature. *Child Development*, *77*(5), 1282-1297. doi: doi.org/10.1111/j.1467-8624.2006.00934.x

Harwood, R. L., Handwerker, W. P., Schoelmerich, A., & Leyendecker, B. (2001). Ethnic category labels, parental beliefs, and the contextualized individual: An exploration of the individualism-sociocentrism debate. *Parenting: Science and Practice*, *1*(3), 217-236. doi: 10.1207/ S15327922PAR0103_03

Harwood, R. L., Schoelmerich, A., Ventura‐Cook, E., Schulze, P. A., & Wilson, S. P. (1996). Culture and class influences on Anglo and Puerto Rican Mothers' beliefs regarding long‐term socialization goals and child behavior. *Child Development*, *67*(5), 2446-2461. doi: 10.2307/1131633

Haskett, M. E., Ahern, L. S., Ward, C. S., & Allaire, J. C. (2006). Factor structure and validity of the Parenting Stress Index-Short Form. *Journal of Clinical Child & Adolescent Psychology*, *35*(2), 302-312. doi: 10.1207/s15374424jccp3502_14

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9*(2), 139-164. doi: 10.1177/014662168500900204

Haynes, S. N., Nelson, K., & Blaine, D. D. (1999). Psychometric issues in assessment research. In P. C. Kendall, J. N. Butcher, & G. N. Holmbeck (Eds.), *Handbook of research methods in clinical psychology, second edition* (pp. 125-154). Hoboken, NJ: Wiley.

Henggeler, S.W., & Lee, T. (2003). Multisystemic treatment of serious clinical problems. In A.E. Kazdin & J.R. Weisz (Eds.), *Evidence-based psychotherapies for children and adolescents* (pp. 301-322). NewYork: Guilford.

Hillemeir, M. M., Foster, M., Heinrichs, B., & Heier, B. (2007). Racial differences in parental reports of attention-deficit/ hyperactivity disorder behaviors. *Journal of Developmental & Behavioral Pediatrics, 28*, 353-361. doi: 10.1097/DBP.0b013e31811ff8b8

Hinshaw, S. P. (1987). On the distinction between attentional deficits/hyperactivity and conduct problems/aggression in child psychopathology. *Psychological Bulletin*, *101*(3), 443. doi: 10.1037/0033-2909.101.3.443

Hogan, T. P., Benjamin, A., & Brezinksi, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement*, 60, 523–532.

Hollan, D. (1992). Cross-cultural differences in the self. *Journal of Anthropological Research*, *48*(4), 283-300. doi: 10.1086/jar.48.4.3630440

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55. doi: 10.1080/10705519909540118

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology*, *20*(3), 296-309. doi: 10.1177/0022022189203004

Hukkelberg, S. S. (2017). A Reexamination of child problem behaviors as measured by ECBI: Factor structure and measurement invariance across two parent training interventions. *Assessment*. Advance online publication. doi:10.1177/1073191117706022

Hulbert, T. A., Gdowski, C. L., & Lachar, D. (1986). Interparent agreement on the Personality Inventory for Children: Are substantial correlations sufficient? *Journal of Abnormal Child Psychology*, *14*(1), 115-122.

Hunsley, J., & Mash, E. J. (2007). Evidence-based assessment. *Annual Review of Clinical Psycholo*gy *3*, 29-51. doi: 10.1146/annurev.clinpsy.3.022806.091419

IBM Corp. (2017). IBM SPSS Statistics for Windows (Version 25) [Computer Software]. Armonk, NY: IBM Corp.

Ismaili, E. (2014). Psychometric properties of Eyberg Child Behavior Inventory in Albanian context. *International Journal of Academic Research in Progressive Education and Development*, *4*(1), 118-130. doi: 10.6007/IJARPED/v4-i1/1617

Jeter, K., Zlomke, K., Shawler, P., & Sullivan, M. (2017). Comprehensive psychometric analysis of the Eyberg Child Behavior Inventory in children with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, *47*(5), 1354-1368. doi: 10.1007/s10803-017-3048-x

Kamphaus, R. W., Petoskey, M. D., & Rowe, E. W. (2000). Current trends in psychological testing of children. *Professional Psychology: Research and Practice*, *31*(2), 155-164. doi: 10.1037/0735-7028.31.2.155

Kataoka, S. H., Zhang, L., & Wells, K. B. (2002). Unmet need for mental health care among US children: Variation by ethnicity and insurance status. *American*

Journal of Psychiatry, *159*(9), 1548-1555. Retrieved from https://search-proquest-com.ezproxylocal.library.nova.edu /docview/220477722?accountid=6579

Kazdin, A. E. (2003). Problem-solving skills training and parent management training for oppositional defiant disorder and conduct disorder. In A.E. Kazdin, & J.R. Weisz (Eds.), *Evidence-based psychotherapies for children and adolescents* (pp. 211-226).

Kazdin, A. E. (2005). Evidence-based assessment for children and adolescents: Issues in measurement development and clinical application. *Journal of Clinical Child and Adolescent Psychology*, *34*(3), 548-558. doi: 10.1207/s15374424jccp3403_10

Kean, J., & Reilly, J. (2014). Item response theory. In F.M. Hammond, J.F. Malec, T.G. Nick, & R.M. Buschbacher (Eds.), *Handbook for clinical research: Design, statistics and implementation* (pp. 195-198). Retrieved from: https://static1.squarespace.com/static/514fd024e4b0d4d5c3e59e38/t/53bc6115e4b07f64b249600c/1404854549244/Kean+Reilly+%282014+in+press%29+Item+Response+Theory.pdf

Kolko, D. J., & Kazdin, A. E. (1993). Emotional/behavioral problems in clinic and nonclinic children: Correspondence among child, parent and teacher reports. *Journal of Child Psychology and Psychiatry*, *34*(6), 991-1006. doi: 10.1111/j.1469-7610.1993.tb01103.x

Langer, M. M., Hill, C. D., Thissen, D., Burwinkle, T. M., Varni, J. W., & DeWalt, D. A. (2008). Item response theory detected differential item functioning between healthy and ill children in quality-of-life measures. *Journal of Clinical Epidemiology*, *61*(3), 268-276.

Lawton, K. E., & Gerdes, A. C. (2014). Acculturation and Latino adolescent mental health: Integration of individual, environmental, and family influences. *Clinical Child and Family Psychology Review*, *17*(4), 385-398. doi: 10.1007/s10567-014-0168-0

Leung, C., Sanders, M. R., Leung, S., Mak, R., & Lau, J. (2003). An outcome evaluation of the implementation of the triple P‐Positive Parenting Program in Hong Kong. *Family Process*, *42*(4), 531-544. doi: 10.1111/j.1545-5300.2003.00531.x

Levitt, J. M., Saka, N., Romanelli, L. H., & Hoagwood, K. (2007). Early identification of mental health problems in schools: The status of instrumentation. *Journal of School Psychology*, *45*(2), 163-191. doi: 10.1016/j.jsp.2006.11.005

Linacre, J. M. (1995). Misfit Statistics for Rating Scale Categories. *Rasch Measurement Transactions, 9(*3*),* 450. Retrieved from https://www.rasch.org/rmt/rmt93j.htm

Linacre, J. M. (1998). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement*, *2*, 266-283. Retrieved from https://pdfs.semanticscholar.org/b4a8/0ea0deb226c438a5f74b7f7f66a9a4b587c4.pdf

Linacre, J. M. (1999). Category disordering (disordered categories) vs. threshold disordering (disordered thresholds). *Rasch Measurement Transactions, 13*(1)*,* 675. Retrieved from https://www.rasch.org/rmt/rmt131a.htm

Linacre, J. M. (2000). Comparing and choosing between "Partial Credit Models" (PCM) and "Rating Scale Models" (RSM*). Rasch Measurement Transactions, 14(3),* 78. Retrieved from https://www.rasch.org/rmt/rmt143k.htm

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*(1), 85-106. doi: 10.1.1.424.2811

Linacre, J. M. (2010). A user's guide to WINSTEPS/MINISTEP: Rasch-model computer programs. Program manual 3.93.0. Retrieved from http://www.winsteps.com/winman/

Linacre, J.M. (2017). Winsteps® (Version 4.0.0) [Computer Software]. Beaverton, Oregon: Winsteps.com. Retrieved December 2, 2017. Available from https://www.winsteps.com/.

Linacre, J. M. (2017). Winsteps® Rasch measurement computer program User's Guide. Beaverton, Oregon: Winsteps.com

Macy, M. (2012). The evidence behind developmental screening instruments. *Infants & Young Children*, *25*(1), 19-61. doi: 10.1097/IYC.0b013e31823d37dd

Martinez, C. R., & Eddy, J. M. (2005). Effects of culturally adapted parent management training on Latino youth behavioral health outcomes. *Journal of Consulting and Clinical Psychology*, *73*(5), 841- 851. doi:10.1037/0022-006X.73.5.841

Mascendaro, P. M., Herman, K. C., & Webster-Stratton, C. (2012). Parent discrepancies in ratings of young children's co-occurring internalizing symptoms. *School Psychology Quarterly, 27*(3), 134-143. doi:10.1037/a0029320

Mash, E.J., & Hunsley, J. (2005) Evidence-based assessment of child and adolescent disorders: Issues and challenges. *Journal of Clinical Child & Adolescent Psychology*, *34*(3), 362-379. doi: 10.1207/s15374424jccp3403_1

Mash, E. J., & Hunsely, J, (2007). Assessment of child and family disturbance: A developmental systems approach. In E.J. Mash, & R.A. Barkley (Eds.),

Assessment of childhood disorders, fourth edition (pp. 3-50). New York, New York: The Guilford Press. Retrieved from: http://file.zums.ac.ir/ebook/061-Assessment%20of %20Childhood%20Disorders,%204th%20Edition-Eric%20J.%20Mash%20PhD %20Russell%20A.%20Barkley%20PhD%20ABPP%20A.pdf#page=16

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174. doi: 10.1007/BF02296272

Matos, M., Torres, R., Santiago, R., Jurado, M., & Rodríguez, I. X. A. (2006). Adaptation of Parent–Child Interaction Therapy for Puerto Rican families: A preliminary study. *Family Process*, *45*(2), 205-222. doi: 10.1111/j.1545-5300.2006.00091.x

Maydeu-Olivares, A., & Montaño, R. (2013). How should we assess the fit of Rasch-type models? Approximating the power of goodness-of-fit statistics in categorical data analysis. *Psychometrika*, *78*(1), 116-133. doi: 10.1007/s11336-012-9293-1

McCabe, K., & Yeh, M. (2009). Parent–Child Interaction Therapy for Mexican Americans: A randomized clinical trial. *Journal of Clinical Child & Adolescent Psychology*, *38*(5), 753-759. doi: 10.1080/15374410903103544

McCabe, K. M., Yeh, M., Garland, A. F., Lau, A. S., & Chavez, G. (2005). The GANA program: A tailoring approach to adapting parent child interaction therapy for Mexican Americans. *Education and Treatment of Children 28*(2), 111-129. Retrieved from http://www.jstor.org. ezproxylocal.library.nova.edu/stable/42899836

Merikangas, K. R., Nakamura, E. F., & Kessler, R. C. (2009). Epidemiology of mental disorders in children and adolescents. *Dialogues in Clinical Neuroscience*, *11*(1), 7-20. Retrieved from https://www-ncbi-nlm-nih gov.ezproxylocal.library .nova.edu/pmc/articles PMC2807642/pdf/DialoguesClinNeurosci-11-7.pdf

Monzó, L. D., & Rueda, R. (2006). A sociocultural perspective on acculturation: Latino immigrant families negotiating diverse discipline practices. *Education and Urban Society*, *38*(2), 188-203. doi: 10.1177/0013124505284293

Morgan, P. L., Staff, J., Hillemeier, M. M., Farkas, G., & Maczuga, S. (2013). Racial and ethnic disparities in ADHD diagnosis from kindergarten to eighth grade. *Pediatrics*, *132*(1), 85-93. doi:10.1542/peds.2012-2390

Murray, A. L., & Johnson, W. (2013). The limitations of model fit in comparing the bi-factor versus higher-order models of human cognitive ability structure. *Intelligence*, *41*(5), 407-422. doi:10.1016/j.intell.2013.06.004

National Institute on Minority Health and Health Disparities (2010). Retrieved from https://www.nimhd.nih.gov.

Niec, L. N., Acevedo-Polakovich, I. D., Abbenante-Honold, E., Christian, A. S., Barnett, M. L., Aguilar, G., & Peer, S. O. (2014). Working together to solve disparities: Latina/o parents' contributions to the adaptation of a preventive intervention for childhood conduct problems. *Psychological Services*, *11*(4), 410-420. doi:10.1037/a0036200.

Nixon, R. D., Sweeney, L., Erickson, D. B., & Touyz, S. W. (2003). Parent-child interaction therapy: A comparison of standard and abbreviated treatments for

oppositional defiant preschoolers. *Journal of Consulting and Clinical Psychology*, *71*(2), 251-260. doi: 10.1037/0022-006X.71.2.251

Olfson, M., Mojtabai, R., Sampson, N. A., Hwang, I., Druss, B., & Kessler, R. C. (2009). Dropout from outpatient mental health care in the United States. *Psychiatric Services*, *60*(7), 898-907. doi: 10.1176/appi.ps.60.7.898

Padilla, A. M., & Medina, A. (2001). Issues in culturally appropriate assessment. In L. Suzuki, J. Ponterotto, & P. Meller (Eds.), *Handbook of multicultural assessment: Clinical, psychological, and educational applications* (pp. 6-21) Retrieved from https://www.researchgate.net/publication/232442877_Issues_in_culturally_appropriate_assessment

Pallant, J. F., Miller, R. L., & Tennant, A. (2006). Evaluation of the Edinburgh Post Natal Depression Scale using Rasch analysis. *BMC Psychiatry*, *6*, 28. doi: 10.1186/1471-244X-6-28

Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology*, *46*(1), 1-18. doi:10.1348/014466506X96931

Pelham, W. E., & Fabiano, G. A. (2008). Evidence-based psychosocial treatments for attention-deficit/hyperactivity disorder. *Journal of Clinical Child & Adolescent Psychology*, *37*(1), 184-214. doi: 10.1080/15374410701818681

Pelham, W.E., Fabiano, G.A., Gnagy, E.M., Greiner, A.R., & Hoza, B. (2005). The role of summer treatment program in the context of comprehensive treatment for Attention Deficit/Hyperactivity Disorder. In E. Hibbs, & P. Jensen (Eds.), *Psychosocial treatments for child and adolescent disorders: Empirically based*

strategies for clinical practice, second edition (pp. 377-410). Washington, D.C.: APA Press

Pelham, W.E., & Hoza, B. (1996). Intensive treatment: A summer treatment program for children with ADHD. In E. Hibbs, & P. Jensen (Eds.), *Psychosocial treatments for child and adolescent disorders: Empirically based strategies for clinical practice* (pp. 311–340). New York: APA Press.

Pew Hispanic Center. (2015). *Statistical portrait of Hispanics in the United States, 2015*. Retrieved from http://www.pewhispanic.org/2017/09/18/facts-on-u-s-latinos/

Phares, V. (1992). Where's poppa? The relative lack of attention to the role of fathers in child and adolescent psychopathology. *American Psychologist*, *47*(5), 656-664. doi: 10.1037/0003-066X.47.5.656

Phares, V. (1997). Accuracy of informants: Do parents think that mother knows best?. *Journal of Abnormal Child Psychology*, *25*(2), 165-171. doi: 10.1023/A:1025787613984

Phares, V., Lopez, E., Fields, S., Kamboukos, D., & Duhig, A. M. (2005). Are fathers involved in pediatric psychology research and treatment?. *Journal of Pediatric Psychology*, *30*(8), 631-643. doi: 10.1093/jpepsy/jsi050

Pigott, R. L., & Cowen, E. L. (2000). Teacher race, child race, racial congruence, and teacher ratings of children's school adjustment. *Journal of School Psychology*, *38*(2), 177-195. doi: 10.1016/S0022-4405(99)00041-2

Plante, T. G., Couchman, C. E., & Diaz, A. R. (1995). Measuring treatment outcome and client satisfaction among children and families. *The Journal of Behavioral Health Services and Research*, *22*(3), 261-269. doi: 10.1037/0735-7028.29.1.52

Poston, J. M., & Hanson, W. E. (2010). Meta-analysis of psychological assessment as a therapeutic intervention. *Psychological Assessment*, 22, 203–212. doi:10.1037/a0018679

Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*, *2*(1),13-43. doi:10.1207/S15328031US0201_02

Pumariega, A. J., Rogers, K., & Rothe, E. (2005). Culturally competent systems of care for children's mental health: Advances and challenges. *Community Mental Health Journal*, *41*(5), 539-555. doi: 0.1007/s10597-005-6360-4

Raîche, G. (2005). Critical eigenvalue sizes (variances) in standardized residual principal components analysis. *Rasch Measurement Transactions 19*(1) 1012. Retrieved from http://www.rasch.org/rmt/rmt191h.htm

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Chicago: MESA Press.

Reedtz, C., Bertelsen, B., Lurie, J. I. M., Handegård, B. H., Clifford, G., & Mørch, W. T. (2008). Eyberg Child Behavior Inventory (ECBI): Norwegian norms to identify conduct problems in children. *Scandinavian Journal of Psychology*, *49*(1), 31-38. doi: 10.1111/j.1467-9450.2007.00621.x

Reid, M.J., Webster-Stratton, C., & Hammond, M. (2003). Follow-up of children who received the Incredible Years Intervention for Oppositional-Defiant Disorder: Maintenance and prediction of 2-year outcome. *Behavior Therapy 34* (4), 471-491. doi: 10.1016/S0005-7894(03)80031

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667-696. doi: 10.1080/00273171.2012.715555

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment*, *95*(2), 129-140. doi: 10.1080/00223891.2012.725437

Rescorla, L. A., Ginzburg, S., Achenbach, T. M., Ivanova, M. Y., Almqvist, F., Begovac, I., ... & Döpfner, M. (2013). Cross-informant agreement between parent-reported and adolescent self-reported problems in 25 societies. *Journal of Clinical Child & Adolescent Psychology*, *42*(2), 262-273. doi: 10.1080/15374416.2012.717870

Rhee, E. R., & Rhee, E. S. (2015). Psychometric properties and standardization of the Korean version of the Eyberg Child Behavior Inventory. *Journal of Child and Family Studies*, *24*(8), 2453-2462. doi: 10.1007/s10826-014-0048-8

Rich, B. A., & Eyberg, S. M. (2001). Accuracy of assessment: The discriminative and predictive power of the Eyberg Child Behavior Inventory. *Ambulatory Child Health*, *7*(3‐4), 249-257. Retrieved from http://web.a.ebscohost.com.ezproxy local.library.nova.edu/ehost/pdfviewer/pdfviewer?vid=1&sid=5cd62f3f-2ec1-43d5-9a86-c51622171863%40sessionmgr4007

Richters, J. E. (1992). Depressed mothers as informants about their children: A critical review of the evidence for distortion. *Psychological Bulletin*, *112*, 485-485. doi:10.1037/0033-2909.112.3.485

Ringel, J. S., & Sturm, R. (2001). National estimates of mental health utilization and expenditures for children in 1998. *The Journal of Behavioral Health Services and Research*, *28*(3), 319-333. doi: 10.1007/BF02287247

Robinson, E. A., Eyberg, S. M., & Ross, A. W. (1980). The standardization of an inventory of child conduct problem behaviors. *Journal of Clinical Child & Adolescent Psychology*, *9*(1), 22-28. doi:10.1080/15374418009532938

Rosselló, J., & Bernal, G. (1999). The efficacy of cognitive-behavioral and interpersonal treatments for depression in Puerto Rican adolescents. *Journal of Consulting and Clinical Psychology*, *67*(5), 734-745. doi:10.1037/0022-006X.67.5.734

Rothe, E. M. (2005). Considering cultural diversity in the management of ADHD in Hispanic patients. *Journal of the National Medical Association*, *97*(10) Supplement,175-235. Retrieved from https://search-proquest-com.ezproxylocal.library.nova.edu/ docview/214054762?pq-origsite=gscholar

Rush, A. J., Gullion, C. M., Basco, M. R., Jarrett, R. B., & Trivedi, M. H. (1996). The Inventory of Depressive Symptomatology (IDS): Psychometric properties. *Psychological Medicine*, *26*(3), 477-486. doi: 10.1017/S0033291700035558

Sabogal, F., Marín, G., Otero-Sabogal, R., Marín, B. V., & Perez-Stable, E. J. (1987). Hispanic familism and acculturation: What changes and what doesn't? *Hispanic Journal of Behavioral Sciences*, *9*(4), 397-412. doi:10.1177/07399863870094003

Salbach-Andrae, H., Lenz, K., & Lehmkuhl, U. (2009). Patterns of agreement among parent, teacher and youth ratings in a referred sample. *European Psychiatry*, *24*(5), 345-351. doi:10.1016/j.eurpsy.2008.07.008

Sawyer, M. G., Baghurst, P., & Clark, J. (1992). Differences between reports from children, parents and teachers: Implications for epidemiological studies. *Australian & New Zealand Journal of Psychiatry*, *26*(4), 652-660. doi: 10.3109/00048679209072102

Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment*, *29*(4), 304-321. doi:10.1177/0734282911406653

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*(4), 350-353. doi: 10.1037/1040-3590.8.4.350

Schmitz, M. F., & Velez, M. (2003). Latino cultural differences in maternal assessments of attention deficit/hyperactivity symptoms in children. *Hispanic Journal of Behavioral Sciences*, *25*(1), 110-122. doi: 10.1177/0739986303251700

Schneider, H., & Eisenberg, D. (2006). Who receives a diagnosis of attention-deficit/hyperactivity disorder in the United States elementary school population?. *Pediatrics*, *117*(4), e601-e609. doi: 10.1542/peds.2005-1308

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., Graeff, A. D., Groenvold, M., … Sprangers, M. A. G. (2009). A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales. *Journal of Clinical Epidemiology, 62,* 288- 295. doi: 10.1016/j.jclinepi.2008.06.003

Shernoff, E. S., Hill, C., Danis, B., Leventhal, B. L., & Wakschlag, L. S. (2014). Integrative consensus: A systematic approach to integrating comprehensive

assessment data for young children with behavior problems. *Infants & Young Children*, 27(2), 92-110. doi:10.1097/IYC.0000000000000008

Shweder, R. A., & Bourne, E. J. (1982). Does the concept of the person vary cross-culturally?. In A.J Marsella, & G.M. White (Eds.), *Culture, illness, and healing* (pp. 97–137). Netherlands: Springer. doi:10.1007/978-94-010-9220-3

Sivan, A. B., Ridge, A., Gross, D., Richardson, R., & Cowell, J. (2008). Analysis of two measures of child behavior problems by African American, Latino, and Non-Hispanic European American parents of young children: A focus group study. *Journal of Pediatric Nursing*, *23*(1), 20-27. doi:10.1016/j.pedn.2007.07.005

Slopen, N., Fitzmaurice, G., Williams, D. R., & Gilman, S. E. (2010). Poverty, food insecurity, and the behavior for childhood internalizing and externalizing disorders. *Journal of the American Academy of Child & Adolescent Psychiatry*, *49*(5), 444-452. doi: 10.1016/j.jaac.2010.01.018

Smith, R. M. (1991). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement*, *51*(3), 541-565. doi: 10.1177/0013164491513003

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology*, *8*(1), 33. doi:10.1186/1471-2288-8-33

Smith, R. M., Schumacker, R. E., & Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement*, *2*(1), 66. Retrieved from http://files.eric.ed.gov.ezproxylocal.library.nova.edu/fulltext/ED384617.pdf

Stern, M. (2007). A*ssessing dimensions of disruptive child behavior with the Eyberg Child Behavior Inventory*. Unpublished manuscript, Department of Psychology, University of Florida, Gainesville, Florida. Retrieved from http://ufdcimages.uflib.ufl.edu/uf/e0/ 02/01/22/00001/stern_m.pdf

Suárez‐Falcón, J. C., & Glas, C. A. (2003). Evaluation of global testing procedures for item fit to the Rasch model. *British Journal of Mathematical and Statistical Psychology*, *56*(1), 127-143. Retrieved from http://web.b.ebscohost.com.ezproxylocal.library.nova.edu /ehost/pdfviewer/pdfviewer?vid=1&sid=1ff8e5b6-2dd3-42c3-b538-51648dc1e2a5%40sessionmgr102

Tennant, A., McKenna, S. P., & Hagell, P. (2004). Application of Rasch analysis in the development and application of quality of life instruments. *Value in Health*, *7*, S22-S26. doi: 10.1111/j.1524-4733.2004.7s106.x

Tennant, A., & Pallant, J. F. (2006). Unidimensionality matters. *Rasch Measurement Transactions*, *20*(1), 1048-1051. Retrieved from https://www.rasch.org/rmt/rmt201c.htm

Tennant, A., Penta, M., Tesio, L., Grimby, G., Thonnard, J. L., Slade, A., ... & Tripolski, M. (2004). Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: The PRO-ESOR project. *Medical Care*, *42*(1), I-37. doi: 10.1097/01.mlr. 0000103529.63132.77

Thomas, M. L. (2011). The value of item response theory in clinical assessment: A review. *Assessment*, *18*(3), 291-307. doi: 10.1177/1073191110374797

Treutler, C. M., & Epkins, C. C. (2003). Are discrepancies among child, mother, and father reports on children's behavior related to parents' psychological symptoms and aspects of parent–child relationships?. *Journal of Abnormal Child Psychology*, *31*(1), 13-27. doi: 10.1023/A:1021765114434

U.S. Census Bureau (2016). *Population Quickfacts 2010-2016: Florida.* Retrieved from https://www.census.gov/quickfacts/fact/table/miamidadecountyflorida,browardcountyflorida,FL/PST045216

U.S. Department of Health and Human Services, Health Resources and Services Administration, Maternal and Child Health Bureau. (2010). *The National Survey of Children's Health 2007.* Retrieved from https://mchb.hrsa.gov/nsch/07emohealth/national/mhs/ pages/2mhs.html

U.S. Public Health Service. (2000). *Report of the Surgeon General's Conference on Children's Mental Health: A National Action Agenda.* Retrieved from https://www.ncbi.nlm.nih.gov /books/NBK44233/

U.S. Surgeon General. (2001). *Mental health: Culture, race, and ethnicity. A supplement to mental health: A report of the surgeon general*. Rockville, MD: U.S.

Department of Health and Human Services. Retrieved from https://www-ncbi-nlm-nih-gov.ezproxylocal.library.nova.edu/books/NBK44243/

Van de Vijver, F. J. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment, 13*, 29-37. doi: 10.1027/1015-5759.13.1.29

Van de Vijver, F. J., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R.K. Hambleton, P.F. Merenda, & C.D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp.39-64) New Jersey: Lawrence Erlbaum Associates. doi:10.4324/9781410611758

Van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology*, *54*(2), 119-135. doi:10.1016/j.erap.2003.12.004

Van de Oord, S., Prins, P., Oosterlaan, J., & Emmelkamp, P. (2006). The association between parenting stress, depressed mood and informant agreement in ADHD and ODD. *Behaviour Research Therapy, 44*(11), 1585-1595. doi: 10.1016/j.brat.2005.11.011

Verhelst, N. D., & Glas, C. A. (1995). The one parameter logistic model. In G.H. Fischer, & I.W. Molenaar (Eds), *Rasch models* (215-238). New York: Springer. doi: 978-1-4612-4230-7

Visser, S. N., Danielson, M. L., Bitsko, R. H., Holbrook, J. R., Kogan, M. D., Ghandour, R. M., ... & Blumberg, S. J. (2014). Trends in the parent-report of health care

provider-diagnosed and medicated attention-deficit/hyperactivity disorder: United States, 2003–2011. *Journal of the American Academy of Child & Adolescent Psychiatry*, *53*(1), 34-46. doi: 10.1016/j.jaac.2013.09.001

Visser, S. N., Zablotsky, B., Holbrook, J. R., Danielson, M. L., & Bitsko, R. H. (2015). Diagnostic experiences of children with attention-deficit/hyperactivity disorder. *National Health Statistics Reports*, (81), 1-7. Retrieved from https://www.cdc.gov/nchs/data/ nhsr/nhsr081.pdf

Von Stauffenberg, C., & Campbell, S. B. (2007). Predicting the early developmental course of symptoms of attention deficit hyperactivity disorder. *Journal of Applied Developmental Psychology*, *28*(5), 536-552. doi: 0.1016/j.appdev.2007.06.011

Walker, C. M. (2011). What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. *Journal of Psychoeducational Assessment*, *29*(4), 364-376. doi: 10.1177/0734282911406666

Wardenaar, K. J., van Veen, T., Giltay, E. J., den Hollander-Gijsman, M., Penninx, B. W., & Zitman, F. G. (2010). The structure and dimensionality of the Inventory of Depressive Symptomatology Self Report (IDS-SR) in patients with depressive disorders and healthy controls. *Journal of Affective Disorders*, *125*(1), 146-154. doi: 10.1016/j.jad.2009.12.020

Webster-Stratton, C., & Hammond, M. (1997). Treating children with early-onset conduct problems: A comparison of child and parent training interventions. *Journal of Consulting and Clinical Psychology*, *65*(1), 93-109. doi:10.1037//0022-006X.65.1.93

Weis, R., Lovejoy, M. C., & Lundahl, B. W. (2005). Factor structure and discriminative validity of the Eyberg Child Behavior Inventory with young children. *Journal of Psychopathology and Behavioral Assessment*, 27(4), 269-278. doi: 10.1007/s10862-005-2407-7

Weitzman, C., & Wegner, L. (2015). Promoting optimal development: Screening for behavioral and emotional problems. *Pediatrics*, 133(2), 384-395. doi: 10.1542/peds.2014-3716

Whiteside-Mansell, L., Ayoub, C., McKelvey, L., Faldowski, R. A., Hart, A., & Shears, J. (2007). Parenting stress of low-income parents of toddlers and preschoolers: Psychometric properties of a short form of the Parenting Stress Index. *Parenting: Science and Practice*, *7*(1), 26-56.

Wright, B. D. (1996). Comparing Rasch measurement and factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *3*(1), 3-24. doi: 10.1080/10705519609540026

Wright, B.D. (1996) Local dependency, correlations and principal components. *Rasch Measurement Transactions,* 10, 3, 509-511. Retrieved from https://www.rasch.org/rmt/rmt103b.htm

Youngstrom, E., Loeber, R., & Stouthamer-Loeber, M. (2000). Patterns and correlates of agreement between parent, teacher, and male adolescent ratings of externalizing and internalizing problems. *Journal of Consulting and Clinical Psychology*, *68*(6), 1038-1050. doi: 10.1037/0022-006X.68.6.1038

Yu, C. H., Popp, S. O., DiGangi, S., & Jannasch-Pennell, A. (2007). Assessing unidimensionality: A comparison of Rasch modeling, parallel analysis, and TETRAD. *Practical Assessment, Research & Evaluation*, *12*(14). Retrieved from https://www.researchgate.net/publication/241422758_Assessing_unidimensionality_A_comparison_of_Rasch_Modeling_Parallel_Analysis_and_TETRAD

Yabiku, S., Kulis, S., Marsiglia, F., Lewin, B., Nieri, T., & Hussaini, S. (2007) Neighborhood effects on the efficacy of a program to prevent youth alcohol use. *Substance Use and Misuse 42(*1), 65-87. doi: 10.1080/10826080601094264

Zaidman‐Zait, A., Mirenda, P., Zumbo, B. D., Wellington, S., Dua, V., & Kalynchuk, K. (2010). An item response theory analysis of the Parenting Stress Index‐Short Form with parents of children with autism spectrum disorders. *Journal of Child Psychology and Psychiatry*, *51*(11), 1269-1277. doi: 10.1111/j.1469-7610.2010.02266.x

Appendix A

CFA of the Five-Point Rating Scale Structure

In order to confirm the findings regarding the superior fit of the three-factor model (hypothesis one) for the five-point rating scale structure, CFA was performed with the rescored data for the 22-items of the ECBI. Overall, the results of the one- and three-factor CFA were similar to those of the seven-point rating scale structure. The three-factor model with three correlated errors provided an acceptable fit, χ^2 (203) = 371.16, $p < .0001$. The CFI, SRMR, and RMSEA values were 0.932, 0.060, and 0.061, respectively. The one-factor model with three correlated errors resulted in a poor fit, χ^2 (206) = 834.412, $p < .0001$. The CFI, SRMR, and RMSEA values were 0.748, 0.104, and 0.118, respectively. In addition to the fit indices reported, the chi-square difference test indicated that the three-factor model with three correlated errors provided a significantly better fit than the one-factor model with three correlated errors, $\Delta\chi^2$ (3) = 463.252, $p <$ .0001. Additionally, the Akaike Information Criterion (AIC) for the three-factor model with three correlated errors resulted in the smallest value across all the models tested (AIC = 471.16), indicating that this model is the best fit for the data. The results of the CFA of the one- and the three-factor models with one, two, and three correlated error terms using the five-point rating scale structure are summarized in Table A1.

Table A1

Model Fit Indices of the One- and Three-Factor Models with Varying Correlated Error Terms for the Five-point Rating Scale Structure

Model	χ^2	df	χ^2 /df	CFI	SRMR	RMSEA	AIC
No Correlated Errors							
1-Factor	1045.518*	209	5.002	0.664	0.1142	0.137	1133.518
3-Factor	556.655*	206	2.702	0.859	0.0711	0.089	650.655
One Correlated Error							
1-Factor	928.259*	208	4.463	0.722	0.1085	0.125	1018.259
3-Factor	424.968*	205	2.073	0.912	0.0603	0.070	520.968
Two Correlated Errors							
1-Factor	859.551*	207	4.152	0.738	0.1055	0.120	951.551
3-Factor	396.385*	204	1.943	0.923	0.0607	0.065	494.385
Three Correlated Errors							
1-Factor	834.412*	206	4.051	0.748	0.1049	0.118	928.412
3-Factor	371.16*	203	1.828	0.932	0.0602	0.061	471.160

*$p < .001$.

Appendix B

Pearson's Product Moment Correlation Coefficients for the Seven-Point Rating Scale

In order to assess the convergent validity of the ECBI scales using the seven-point rating scale structure, the extent to which the total scores of the ODBTA, CPB, and IB scales were correlated with measures of related constructs was examined. Specifically, it was expected that the ODBTA scale would positively correlate with the Oppositional Defiant Disorder Symptom Scale of the Conners Parent Rating Scale, 3rd Edition (CPRS). The CPB scale was expected to correlate positively with the Conduct Disorder Symptom Scale of the CPRS. Finally, the IB scale would positively correlate with the ADHD Predominately Inattentive Symptom Scale of the CPRS. Furthermore, for the CPRS Content Scales, it was expected that the IB scale would positively correlate with the Inattention Content Scale of the CPRS, given the similar content of the two scales.

In order to assess the discriminant validity of the ECBI scales, the extent to which the scores of the ODBTA, CPB, and IB seven-point scales were correlated with theoretically unrelated variables was examined. It was expected that the ODBTA, CPB, and IB scale scores would not be strongly correlated with either the Peer Relations or the Learning Problems Content Scale scores of the CPRS.

Pearson's product moment correlation coefficients, *r,* was used to examine the relationships between the ODBTA, CPB, and IB seven-point scale scores and the CPRS scale scores. Pearson's *r* is a parametric test used to measure the strength of association between two variables and is appropriate for continuous variables. An *r* value of one indicates a perfect positive correlation, a value of negative one indicates a perfect negative correlation, and a value of zero indicates no correlation. In order to assess the strength of the correlation, the guidelines suggested by Evans (1996) were used (Table 29).

The correlation analyses included 194 of the cases with CPRS data out of the total sample (N = 221). The CPRS is appropriate for parents of children ages six to 18. Of the 27 excluded cases, 21 were excluded due age, i.e., younger than six years of age. The remaining six cases were excluded due to missing CPRS data. The subsample (n = 194) was 72.2% male and 27.8% female, with an average age of 9.8 years (range six – 17, *SD* = 2.67). A total of 67.9% of the parents in this subsample were married; 17.1% were divorced; 2.6% were separated; 9.3% were single and had never been married; 2.1% were living with someone; 1% were widowed; and one respondent's marital status was missing. Regarding ethnicity, 42.3% of the sample was Hispanic, 41.2% was European American, 12.9% was African American, and 3.6% identified as "other." The Pearson correlation coefficients for the relationships between the ODBTA, CPB, and IB seven-point scale scores and the CPRS scale scores are shown in Table B1.

Table B1

Pearson's Product Moment Correlation Coefficients for the Seven-Point Rating Scale

	Pearson's *r*		
	IB	ODBTA	CPB
CPRS Symptom Scales			
ADHD Predominately Inattentive Type	0.530**	0.334**	0.184**
ADHD Predominately Hyperactive-Impulsive Type	0.386**	0.515**	0.412**
Conduct Disorder	0.101	0.567**	0.574**
Oppositional Defiant Disorder	0.299**	0.688**	0.523**
CPRS Content Scales			
Peer Relations	0.168**	0.369**	0.271**
Aggression	0.150*	0.496**	0.495**
Learning Problems	0.214**	0.305**	0.300**
Executive Functioning	0.543**	0.292**	0.143*
Inattention	0.613**	0.297**	0.190**
Hyperactivity/Impulsivity	0.398**	0.500**	0.355**

Note: CPRS= Conners Parent Rating Scale, 3rd Edition*;* ECBI= Eyberg Child Behavior Inventory; IB = Inattentive Behaviors; ODBTA = Oppositional Defiant Behavior Toward Adults; and CPB = Conduct Problem Behavior.
**p-value <.05, ** p-value <.01*

The results in Table B1 show that there was a moderate, positive correlation between the IB scale score and the ADHD Predominately Inattentive Type Symptom scale score, i.e., $r = 0.530$, n = 194, $p < .01$, and a strong, positive correlation between the IB scale score and the Inattention Content Scale score, i.e., $r = 0.613$, n = 194, $p < .01$. Additionally, there was a strong, positive correlation between the ODBTA scale score and the Oppositional Defiant Disorder Symptom scale score, i.e., $r = 0.688$, n = 194, $p < .01$. Finally, there was a moderate, positive correlation between the CPB scale score and the Conduct Disorder Symptom scale score, i.e., $r = 0.574$, n = 194, $p < .01$. These results provide evidence for the convergent validity of the three ECBI scales.

Regarding the discriminant validity of the IB, ODBTA, and CPB scales, the IB ($r = 0.168$, n = 194, $p < .01$); ODBTA ($r = 0.369$, n = 194, $p < .01$); and CPB ($r = 0.271$, n

= 194, $p < .01$) scale scores were weakly correlated with the Peer Relations Content Scale score of the CPRS. Similarly, the IB ($r = 0.214$, n = 194, $p < .01$); ODBTA ($r = 0.305$, n = 194, $p < .01$); and CPB ($r = 0.300$, n = 194, $p < .01$) scale scores were weakly correlated with the Learning Problems Content Scale score of the CPRS. These results provide evidence for the discriminant validity of the three ECBI scales.