Impacts of Factor Analytic Choices on Scale Interpretation Commonly Used to Assess Children With and Without Autism Spectrum Disorder

Abstract

IMPACTS OF FACTOR ANALYTIC CHOICES ON SCALE INTERPRETATION COMMONLY USED TO ASSESS CHILDREN WITH AND WITHOUT AUTISM SPECTRUM DISORDER

By

Abeera Rehmani

Tuesday April 22, 2025

Introduction: Autism Spectrum Disorder (ASD) measures are often designed as diagnostic tools but also used as outcome variables in research. This dual usage poses validation challenges since these tools are typically assessed only for diagnostic accuracy, not as outcome measures. This issue is particularly pronounced for ordinal scales, where their utility as outcome variables is more complex than their validation for clinical diagnostics.

Objective: This study examines how various factor analytic decisions influence the quality of scales, using the Childhood Symptom Behavior Questionnaire (CSBQ) as a model, focusing on its development as an outcome variable.

Methods: Using responses from approximately 7,000 children in the CDC's Pathway's dataset, this study compares four factor analytic methods—Maximum Likelihood (ML), Categorical ML (cML), Diagonally Weighted Least Squares (DWLS), and Unweighted Least Squares (ULS). These methods are evaluated using fit indices in their regular, scale-shifted, and robust forms to determine their impact on model fit. Additionally, we engaged in the questionable research practice (QRP) of correlating item error terms to “fix” item-factor relationships.

Results: Analysis revealed substantial variations in model fit across methods. Scale-shifted and robust conditions notably affected model stability, highlighting the importance of selecting appropriate analytic techniques for ordinal data. Engagement in the QRP of correlating errors to "fix" models was variably required with strongest implications for scale shifted and robust DWLS and ULS models wherein regular unadjusted models did not require fixing but scale-shifted and robust did.

Conclusion: The choice of factor analytic method significantly influences scale interpretation accuracy in ASD assessments and conditions eliciting the QRP of "fixing" error terms poorly outlined. This study advocates using methods that accommodate the ordinal nature of data, enhancing the utility of ASD scales in both clinical and public health settings.

Impacts of Factor Analytic Choices on Scale Interpretation Commonly Used to Assess Children With and Without Autism Spectrum Disorder

By

Abeera Rehmani

M.P.H., GEORGIA STATE UNIVERSITY

A Thesis Submitted to the Graduate Faculty

of Georgia State University in Partial Fulfillment

of the

Requirements for the Degree

MASTER OF PUBLIC HEALTH

ATLANTA, GEORGIA

30303

APPROVAL PAGE

IMPACTS OF FACTOR ANALYTIC CHOICES ON SCALE INTERPRETATION COMMONLY USED TO ASSESS CHILDREN WITH AND WITHOUT AUTISM SPECTRUM DISORDER

By

Abeera Rehmani

Approved:

Brian Barger
Committee Chair

Ike Okosun
Committee Member

April 22, 2025
Date

Dedication

This thesis is dedicated to my parents, whose unwavering support and sacrifices have shaped the person I am today. To my mother and father, who have always worked tirelessly and faced every hardship with resilience, all in the pursuit of a better future for our family. Your endless love, strength, and encouragement inspire me to strive for excellence and to contribute meaningfully to the world. This accomplishment is not only mine but also yours; it is a testament to your dreams, your struggles, and your unyielding hope.

To my siblings and my niece, along with all my other family members—without you, I would be lost. Your constant love, laughter, and support have been my anchor and my compass. Thank you for being my home no matter where I am in the world.

Acknowledgements

I am immensely grateful to my professors and mentors, Dr. Brian Barger and Dr. Ike Okosun, whose guidance and expertise have been invaluable throughout this journey. Dr. Barger, your insights into developmental disabilities have profoundly shaped my approach to research and advocacy. Dr. Okosun, your rigorous approach to epidemiological studies has not only honed my analytical skills but also deepened my understanding of complex public health issues. Your support and mentorship have been crucial in navigating the challenges of this thesis and my academic endeavors. Thank you for believing in my potential and for guiding me with patience and wisdom.

I am also deeply thankful to the staff at the School of Public Health. Your support and assistance have been essential to my research and academic success. The environment you foster for learning and development has been incredibly enriching and has played a crucial role in my achievements. Thank you all for your belief in my potential and for providing the resources and encouragement needed to pursue my goals.

Author's Statement Page

In presenting this thesis as a partial fulfillment of the requirements for an advanced degree from Georgia State University, I agree that the Library of the University shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to quote from, to copy from, or to publish this thesis may be granted by the author or, in his/her absence, by the professor under whose direction it was written, or in his/her absence, by the Associate Dean, School of Public Health. Such quoting, copying, or publishing must be solely for scholarly purposes and will not involve potential financial gain. It is understood that any copying from or publication of this dissertation which involves potential financial gain will not be allowed without written permission of the author.

Abeera Rehmani
Signature of Author

TABLE OF CONTENTS

List of Tables

List of Figures

Introduction

Autism Spectrum Disorder (ASD) represents a significant public health concern due to its complex nature and rising global prevalence (Zeidan et al., 2022). Characterized by diverse symptoms that affect communication, social interaction, and behavior, ASD's spectrum nature means that it manifests uniquely across different individuals, making standardized diagnosis challenging (American Psychological Association, 2013). The complexity of ASD, characterized by a spectrum of symptoms ranging from mild to severe impairments in communication, social interaction, and repetitive behaviors, complicates standardized diagnosis and treatment approaches (Maenner et al., 2023). This variability demands robust, flexible diagnostic tools that can adapt to the diverse needs of the affected population and ensure that interventions are tailored to individual profiles, thereby enhancing the overall quality of life and reducing long-term societal and economic burdens. The early onset and potential lifelong implications of ASD necessitate effective diagnostic tools to ensure timely and accurate interventions, which are crucial for improving affected individuals' quality of life (Randall et al., 2018; Falkmer et al., 2013; Akshoomoff et al., 2006).

ASD identification begins with developmental surveillance by healthcare providers and extending through various levels of screening, referrals, and ultimately, formal diagnosis, guided by frameworks like the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) (Ikeda et al., 2014). Surveillance and monitoring involve routine check-ups where healthcare providers monitor children for early signs of ASD. This initial surveillance is critical as early detection can greatly enhance the efficacy of subsequent interventions. Screening for ASD is stratified into two levels: Level 1 screening casts a wide net during pediatric well-child visits to identify children potentially at risk of developmental disorders (Thabtah & Peebles,

2019; Norris & Lecavalier, 2010). Those who exhibit possible ASD indicators are often subjected to Level 2 screenings, which involve more thorough assessments usually carried out by specialists in developmental disorders. Children flagged during Level 2 screenings are referred for further assessments, employing a mix of diagnostic tools and evaluations by multidisciplinary teams, which confirm or rule out an ASD diagnosis using diagnostic tools (Carbone et al., 2020).

Historically, most autism scale development had a primary identification functional utility for clinicians, but over time scales began to be used as outcome variables, which was not their original intent (Moore et al., 2021). Thus, many ASD scales were not designed with the best methods for their use as outcome variables. This is a problem because the optimal methods for developing measures to predict diagnostic classification (i.e., diagnostic accuracy) are different then the assumptions involved when developing measures for predicting traits (Shim et al., 2019; Casson, 2014; Farmer et al., 2024). Further, while diagnostic accuracy methods are uniform and require binary outcomes, the number of methods for developing trait measures such as outcomes is quite extensive with complicated decisions to make depending on the qualities of scales (Toft et al., 2005; Long & Cheng, 2004).

Considered here is the key method of factor analysis (FA). In trait scale design, initially, exploratory factor analysis (EFA) is utilized in early stages to identify potential latent structures within complex symptom data. Once item to scale relationships are developed, confirmatory factor analysis (CFA) is frequently utilized to display that items relate to factors in a hypothesized manner (Matsunaga, 2010). Over the design of an outcome variable, this methodological evolution highlights a critical shift towards more rigorous statistical validation and model fitting, crucial for developing diagnostic tools that are both scientifically robust and clinically applicable (Tavakol & Wetzel, 2020).

Historically, FA is conducted with Maximum Likelihood (ML) with strict assumptions including continuous normally distributed data. Psychological scales frequently do not meet these assumptions and there is encouragement to engage in more sophisticated approaches that recognize the ordinal and often skewed nature of clinical data. This includes techniques such as Diagonally Weighted Least Squares (DWLS) and Unweighted Least Squares (ULS) (Li C.H, 2016). These techniques are advantageous over traditional Maximum Likelihood (ML) methods because they do not require the stringent assumptions of normality and interval measurement, thereby reducing biases and improving the reliability of scale interpretations (Xia & Yang, 2019). For ML and non-ML approaches, there are also robust refinements allowing researchers to correct estimates in the presence of common data challenges (e.g., heavy skew; heterogeneity of variance) (Grove et al., 2021).

This thesis will conduct an outline of how decisions to use different commonly recommended FA methods could influence the interpretation of the Childhood Symptom Behavior Questionnaire (CSBQ) (de Bildt et al., 2009). CSBQ, which is under development, aims to serve as a versatile screening tool and outcome measure in both clinical and research settings. By investigating multiple analytic conditions, this study will delineate how using different methods will impact decisions during scale development (Hartman et al.,2006).

The outcomes of this research are anticipated to have unique implications for the research ASD scale development and public health. By outlining methodological decisions, this study aims to contribute to the development of better designed outcome measures. In summary, this thesis will address critical gaps in the current understanding regarding best practices in the development of ASD diagnostic tools aiming to be used for both identification and as outcome variables.

Literature Review

Process of Identification for Autism Spectrum Disorders

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental condition characterized by a range of symptoms that affect communication, behavior, and social interactions (Jones & Klin, 2009). The symptoms of autism can vary widely among individuals, but typically include challenges in social reciprocity, difficulties in communication (both verbal and non-verbal), and the presence of restricted or repetitive behaviors (American Academy of Pediatrics, 2006). Other associated features may include sensory sensitivities, anxiety, and co-occurring conditions such as attention-deficit/hyperactivity disorder (ADHD) or intellectual disabilities (Van Steensel, Bögels, & de Bruin, 2013). Factors contributing to the rising prevalence may include improved awareness, enhanced diagnostic criteria, and better access to healthcare services (Hodges et al., 2020).

The identification of autism involves a systematic process that integrates complex cross-system navigation involving community-based surveillance, screening, assessment, and ultimately, diagnosis (Barger et al., 2018). The initial phase, surveillance, typically involves monitoring child development through regular health visits where healthcare providers observe children for any atypical behaviors or developmental delays (Marlow et al., 2019). During well-child visits screening tools are often employed, which can flag potential concerns about autism and prompt further evaluation. If screening yields positive results, a comprehensive assessment is conducted by a qualified professional, which may include a combination of interviews, behavioral observations, and standardized tests that delve into the child's cognitive, social, and language development (Barger et al.,2018; Alpert, 2021). The culmination of this process leads to a formal diagnosis, guided by established criteria in the Diagnostic and Statistical Manual of

Mental Disorders (DSM-5). The identification of ASD, therefore, incorporates a blend of informal surveillance and formal diagnostic processes, emphasizing the necessity of specialized tools to ensure accuracy in assessments (Sheldrick et al., 2019; Brown, 2015; Huerta et al., 2012).

Evolution of ASD Early Identification Tools

The early identification of Autism Spectrum Disorders (ASD) has evolved significantly from early clinical observations and diagnoses to the use of structured instruments capable of identifying children at various developmental stages (*Distinguishing Between Screening and Assessment for Mental and Behavioral Health Problems*, 2014). Historically, the recognition of autism as a distinct neurodevelopmental condition emerged in the mid-20th century, leading to the development of initial diagnostic frameworks such as the Childhood Autism Rating Scale (CARS) in the late 1970s (Simonoff et al., 2008). Initially, the identification processes relied heavily on clinical observations, which, while insightful, often lacked standardization.

Over time, the development of structured tools such as the Autism Diagnostic Interview-Revised (ADI-R) and the Autism Diagnostic Observation Schedule (ADOS) marked a pivotal shift (Lord et al., 1994, 2012). These tools are grounded in rigorous psychometric testing and have been continuously refined to improve reliability and accuracy. This progression from observational techniques to standardized assessments reflects methodological advances that have substantially enhanced the diagnostic precision of autism screenings and helped develop a clinical culture aiming to establish “gold standard” tools (Frigaux et al.,2019).

As clinical assessments were established in relationship to DSM criteria, there was an emergence in the development of high-quality screening tools to aid systems in the early prediction of children ultimately identified with ASD. The M-CHAT, for instance, became

widely promoted in the early 2000s, marking a systematic research agenda to develop a population level screening tool (Dumont-Mathieu & Fein, 2005). This tool has undergone numerous revisions to enhance its sensitivity and specificity, establishing it as one of the most widely implemented tools in clinical practice. These revisions included changes to the scoring criteria, recognizing the importance of developmental milestones in distinguishing between typical and atypical behaviors in toddlers (Robins & Dumont-Mathieu, 2006; Wieckowski et al., 2023). Further, the introduction and adaptation of tools such as the Social Communication Questionnaire (SCQ), available in both Current and Lifetime formats, signifies a trend toward nuanced assessments that integrate a child's developmental history comprehensively. For example, Chesnut et al. (2017) conducted a meta-analysis of the SCQ and found it to be an acceptably accurate screener (AUC = 0.885), though accuracy varied significantly based on methodological decisions such as the use of the Current versus Lifetime versions, age groups screened, and sampling methods. Specifically, the Lifetime version showed substantially higher accuracy than the Current version, and screening children younger than 4 years significantly reduced accuracy. These findings underscore the importance of adhering to validated procedural guidelines when utilizing screening instruments to maximize their predictive capabilities and reliability (Chesnut et al., 2017).

While the primary goal of screening and assessment instruments is to ensure diagnostic accuracy and correct classification, these tools frequently serve a dual purpose as outcome measures in clinical and research settings (Hobart et al., 2007). Historically, this has resulted in tools originally designed for diagnostic purposes being retroactively applied as outcome measures, often without rigorous psychometric validation to support assumptions (Chesnut et al., 2017). A notable exception to this in ASD research is the Social Responsiveness Scale (SRS) was

initially developed with factor analytic approach to characterize subthreshold autism-related behaviors but later became used as an outcome measure. Furthermore, more recently several researchers have noted the challenges of using clinical ASD scales as outcomes underscoring the complexities inherent in using these instruments for dual purposes, particularly addressing both the multi-dimensionality of available scales and that many scales only capture a subset of known traits (Constantino et al., 2004; Uljarević,et al., 2023). Thus, trends indicate a shifting paradigm where both predictive accuracy and outcome utility are evaluated simultaneously during the tool development process, using a blend of sophisticated psychometric approaches (Uljarević et al., 2023). Factor analytic approaches are central to these efforts which help validate and clarify the constructs assessed by these instruments. Consequently, current ASD screening and assessment tools increasingly represent multidimensional evaluations that integrate multiple facets of a child's behavior and development (Chesnut et al., 2017; Constantino et al., 2004).

Interdisciplinary Measurement Cultures in Context of Autism Screening and Assessment

A major task of public health is to translate findings from diverse allied health fields to inform decision making regarding the evidence base of early identification tools (Harris et al., 2001). A major challenge is that there are substantive differences of opinion regarding standards both within and between disciplines. Take, for instance, the measurement traditions of educational and health scale development (Pfeiffer & Blei, 2008). In education, the development of cognitive tests (e.g., IQ) typically begin with traits across the entire population with extreme scores marking potential cases for consideration of “disability” or “gifted” populations (Clements et al., 2020). From this measurement culture, there is frequently a higher premium placed on scale design and refinement from a latent variable from a "purity of measurement" and internal reliability perspective; this may incorporate "diagnostic accuracy" in relation to extreme

groups but tends to focus on implications related to scores as an outcome variable (Revelle, 2024). Notably, during scale development, items are often omitted if not strongly correlated with primary traits to ensure a robust and "pure" measure of a single primary construct of interest (i.e., decrease multi-dimensionality of outcome) (DeVellis & Thorpe, 2021). In contrast, most clinical and health measures place a greater emphasis on screening/assessment predictive validity and diagnostic accuracy of discrete outcomes (i.e., identified cases), with items possibly omitted if deemed poorly predictive of target populations (Harris et al., 2001). In this way, clinical health measurement and assessment deviates from mainstream psychological/educational scale development on the broader population in that screening and assessment tools are primarily evaluative aids of clinical judgement, the ultimate arbiter of a diagnostic category; utility as a scientific outcome measure is often a non-primary, though desired, goal (Hobart et al., 2007; Tavakol & Wetzel, 2020).

In ASD research, screening tools are primarily designed as clinical triaging aids, and assessment tools are typically designed to facilitate clinical judgement, differential diagnosis, and triaging for high-risk clinical populations (Bishop & Lord, 2023). Some tools even come to point of being "deemed" as "good/gold standard" by the field, though evidence does not generally support this for any existing ASD tool (Randall et al., 2018). Eventually, tools are vetted psychometrically regarding their factor structure using latent variable approaches with larger samples. The interpretation of the available screening and assessment tools as outcomes is challenging due to the variety of methods employed across disciplines. Notably, early ASD identification scale development research employs a blend of methodologies from Classical Test Theory (CTT) and latent variable techniques, which are implemented according to various sub-

field norms. As more measures utilize FA during the development phase there is need to better understand this process (Barger, 2024; Uljarević,et al., 2023).

Factor Analysis of Tools Designed for Autism Identification

FA serves as the foundational psychometric method underpinning scale development (DiStefano & Morgan, 2014). Broadly, FA aims to identify underlying latent structures among observed variables, based on parametric assumptions that typically involve normality, linearity, and homoscedasticity. Common stages of factor analytic methods include principal component analysis (PCA), exploratory factor analysis (EFA), and confirmatory factor analysis (CFA). Additionally, methods like exploratory structural equation modeling (ESEM) and invariance testing have emerged but require more extensive discussion beyond the scope here (Statistics Solutions, 2025; Dash & Paul, 2021; Yaşlioğlu & Toplu Yaşlioğlu, 2020; Marsh et al., 2020).

CFA specifically allows researchers to falsify hypotheses about relationships between observed measures and their underlying latent constructs (Yaşlioğlu & Toplu Yaşlioğlu, 2020). Fit indices derived from CFA indicate how well the hypothesized model represents the observed data. This is akin to rules of thumb like p-values $< .05$ with the understanding that indices are helpful decision-making aids, but relatively imprecise and require reasoning by the methodologist. Commonly cited "cut-off values" for fit indices include the Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI), where values greater than 0.95 typically suggest excellent fit, and the Root Mean Square Error of Approximation (RMSEA), where values less than 0.06 indicate good fit (Dash & Paul, 2021).

More broadly, CFA is widely utilized to evaluate the dimensionality of autism identification tools, reflecting the heterogeneous nature of autism symptomatology and challenges of measurement development (Tavakol, M., & Wetzel, 2020). These analyses are

crucial for isolating distinct behavioral dimensions, including social communication deficits, restricted interests, and repetitive behaviors (Li, 2016). A significant challenge in ASD psychometric research arises from the common violation of maximum likelihood (ML) assumptions of continuous normality due to the ordinal nature of most ASD assessment scales. Analysis of ordinal ASD data using ML FA can lead to biased parameter estimates and misinterpretations of the factor structure. To address these challenges, researchers like DiStefano and Lebersfeld advocate for estimation techniques tailored explicitly for ordinal data (DiStefano & Morgan, 2014; Lebersfeld er al., 2021).These approaches include diagonally weighted least squares (DWLS), unweighted least squares (ULS), and ML extensions, like categorical maximum likelihood (cML), which employs ML estimation on polychoric correlation matrices derived from ordinal data. These ordinal-specific methods significantly enhance the accuracy of factor structure estimation and yield more reliable fit indices, thereby improving the validity and interpretability of the tools developed for autism identification. By addressing the complexities associated with ordinal scaling, ordinal estimation techniques like DWLS and ULS facilitate more precise and meaningful interpretations of item-factor relationships (Li, 2016; Brown, 2015).

Beyond the challenges posed by ordinal data, additional complexities in CFA emerge from numerous analytic choices. Adjustments such as scale-shifted (mean-adjusted) and robust (mean and variance adjusted) estimators have been developed to handle violations of assumptions common to clinical data (Li C.H, 2016). These adjustments influence the calculation of standard errors (SEs) and consequently alter fit indices, complicating the interpretation of results. Ultimately, the nuanced selection and interpretation of factor analytic methods, estimators, and fit indices are crucial for accurately modeling ASD measures, thereby ensuring

these instruments effectively capture the multidimensional symptom profiles characteristic of ASD. The current study aims to outline the impact of using different CFA methods on the perception of the quality of an ASD measure via systematically compare the impact of using ML and categorical methods, with and without common adjustments, on the perceived goodness of an ASD screening tool in development (DiStefano & Morgan, 2014; Savalei & Rhemtulla, 2013).

ASD Screening Tool Development Case Study

The Children's Social Behavior Questionnaire (CSBQ) was initially developed by Luteijn et al. (1998) and subsequently refined in 2000, aiming to enhance the behavioral assessment of children with Autism Spectrum Disorders (ASD). This tool was designed to capture a wide array of ASD-related behaviors, spanning from mild to severe presentations. The initial item pool was derived through an extensive literature review, detailed input from caregivers, and clinical observations gathered at the Child and Adolescent Psychiatry Outpatient Clinic in Groningen, the Netherlands. The CSBQ instructs caregivers to rate specific behaviors observed over the preceding two months using a three-point ordinal scale ranging from 'not applicable' to 'often applicable' (Hartman, 2006).

Initial psychometric evaluation of the CSBQ demonstrated strong internal consistency and reliability, affirming its potential utility as a clinical assessment instrument. Subsequent validation efforts indicated substantial inter-rater reliability and adequate test-retest stability; however, it was noted that certain subscales, particularly 'social insight problems,' exhibited somewhat lower reliability indices (Hartman et al., 2006). Convergent validity of the CSBQ was further supported through significant correlations with related measures, such as corresponding subscales of the Child Behavior Checklist (CBCL) and the Autism Behavior Checklist (De Bildt

et al., 2009; De La Roche et al., 2024). These findings indicated that the CSBQ reliably assesses behaviors aligned with broader conceptualizations of ASD. Further refinement of the CSBQ involved a rigorous item selection process guided by expert clinical judgment. Items demonstrating limited relevance to core ASD traits were systematically removed, resulting in a refined instrument containing 68 representative items (Hartman et al., 2006). EFA and CFA analyses played a critical role in verifying the dimensional structure and ensuring the psychometric robustness of the final scale (Hartman et al., 2006; De Bildt et al., 2009).

Given its current developmental status, established yet not widely adopted in clinical practice—the CSBQ provides an ideal candidate for methodological exploration. Thus, the primary purpose of the current study, therefore, is to illustrate the practical implications of employing varied confirmatory factor analytic (CFA) approaches on (a) model acceptance and interpretation, (b) identification of problematic items, and (c) engagement with questionable analytic practices commonly implemented to resolve problematic items and enhance model fit.

Methodology

This study analyzes data available from the CDC's Pathways dataset, which includes comprehensive responses from approximately 7,000 children diagnosed with Autism Spectrum Disorder (ASD), Intellectual Disabilities (ID), or Developmental Delays (DD). To prepare for our analysis, we downloaded the dataset in a SAS file from that we created a CSV file to organize the data and subsequently uploaded it to R for further processing.

Primary Outcome Measure

The primary outcome measure employed in this study is the Children's Social Behavior Questionnaire (CSBQ). The Children's Social Behavior Questionnaire (CSBQ) was selected for this study because it is in a developmental phase with an ultimate goal of population screening for ASD and related developmental challenges. (Hartman et al., 2006). The CSBQ is comprised of six factors aimed at measuring various behavior and social traits in children (**Table 1**). The first factor, *Social Behaviors*, assesses 11 items related to communication difficulties and repetitive actions that reflect challenges in basic social interactions. The second factor, *Emotional Responses*, includes 12 items that evaluate the child's emotional reactivity and anxiety, capturing their sensitivity to emotional stimuli. *Social Integration*, the third factor, comprises 8 items that focus on the child's ability to adapt to social norms and engage with peers, highlighting social adaptability issues. The fourth factor*, Social Communication Skills*, involves 7 items that delve into fundamental social communicative abilities such as initiating conversations, sharing emotions, and interpreting social cues. *Adaptability and Flexibility*, the fifth factor with 8 items, addresses the child's response to change and novelty, crucial for understanding their adaptability in varied situations. Lastly, the sixth factor, *Behavioral Regulation*, consists of 3 items that explore how children regulate their emotions, particularly their management of frustration and

anger, which are critical aspects of behavioral control. Each of these factors collectively provides a comprehensive tool for assessing the breadth of social and behavioral challenges encountered by children with developmental concerns. (Hartman et al., 2006)

Table 1: Factors and their Description

Factor	Name	# Items	Description
F1	*Social Behaviors*	11	Includes items on social behaviors such as communication difficulties and repetitive actions.
F2	*Emotional Responses*	12	Captures emotional responses and sensitivity, including behaviors related to emotional reactivity and anxiety.
F3	*Social Integration*	8	Focuses on challenges in social integration, including difficulties adapting to social norms and interacting with peers.
F4	*Social Communication Skills*	7	Reflects basic social communication skills, such as sharing emotions, initiating conversation, and understanding social cues.
F5	*Adaptability and Flexibility*	8	Addresses behaviors linked to adaptability and flexibility, including resistance to change and novelty.
F6	*Behavioral Regulation*	3	Encompasses behavioral outbursts and regulation, highlighting how children manage frustration and anger.

The initial reliability and validity of the Children's Social Behavior Questionnaire (CSBQ) were established based on its comprehensive form, featuring initially 96 to the final 49 items were designed to evaluate a wide range of behaviors in children with Pervasive Developmental Disorders (PDD) (Hartman et al., 2006). The questionnaire was developed through extensive literature reviews, parental feedback, and clinical insights from the outpatient clinic for child and adolescent psychiatry in Groningen, the Netherlands. Parents rated their child's behaviors from the preceding two months on a three-point scale. Then the items were organized into six subscales, including social behaviors, emotional responses, social integration, social communication skills, adaptability and flexibility, and behavioral regulation (de Bildt et al., 2003). The CSBQ demonstrated strong internal consistency (Cronbach's alpha for all subscales > .75), robust inter-rater reliability (ICC > .70 for all subscales, except anxious/rigid

behaviors, which was low at .64), and solid test-retest reliability. Its convergent validity was affirmed by high correlations with similar subscales of established instruments like the Child Behavior Checklist and the Autism Behavior Checklist at >0.80 except for the subscale ”social insight problem which was at 0.62 and stereotypical at 0.32. This foundational evidence established CSBQ as a reliable and valid tool for detailed assessment of behavioral and social traits in children with PDD (Bildt et al., 2003; Achenbach, 1991; Krug, Arick, & Almond, 1980; Hartman et al., 2006; Brown, 2015).

Statistical Analysis

All statistical analyses were conducted using R. Data were cleaned, recoded, and analyzed to identify, estimate, and evaluate the factor structure and reliability of the Children's Social Behavior Questionnaire (CSBQ) using appropriate packages for psychometric analysis and data visualization. Central to the analysis was Confirmatory Factor Analysis (CFA), a technique used to test whether a hypothesized model of latent constructs aligns with the observed item responses. Conceptually, each questionnaire item was treated as a dependent variable regressed onto one or more latent factors, with factor loadings representing the strength of these relationships. Model estimation was performed in R, with a focus on establishing a valid and interpretable measurement scale for the CSBQ (Marsh et al., 2020; Shim et al., 2019). Internal consistency of each factor was assessed using Cronbach's alpha (α), with $\alpha \geq 0.70$ considered acceptable.

Confirmatory Factor Analysis. CFA was central to this study's approach for validating the factor structure of the Children's Social Behavior Questionnaire (CSBQ). CFA was used to test whether the hypothesized six-factor model—based on prior theoretical and empirical work—appropriately represented the underlying structure of the observed data. In this context, each

questionnaire item was modeled as a dependent variable regressed onto a specific latent factor, with the factor loadings indicating the strength and direction of the relationship between each item and its corresponding construct (Yaşlioğlu & Toplu Yaşlioğlu, 2020). This regression-based framework allowed us to statistically evaluate how well each item measured its intended factor, while accounting for measurement error (de Bildt et al., 2009; Hartman et al.,2006).

A crucial part of our model assessment involved the chi-square statistic, a primary indicator of model fit. This statistic measures the discrepancy between the observed and expected covariance matrices. Ideally, a non-significant chi-square value indicates a model that fits well. However, its effectiveness can be limited by its sensitivity to sample size, which can lead to significant chi-square values even with minor discrepancies in large samples. Consequently, it is often necessary to consider additional fit indices for a more rounded evaluation (Newsom, 2008).

Fit Indices. Model estimation was performed in R using psychometric and structural equation modeling packages. To assess model fit, we evaluated several widely accepted indices, including the Comparative Fit Index (CFI), Root Mean Square Error of Approximation (RMSEA), and Standardized Root Mean Square Residual (SRMR) alongside the chi-square test of model fit. The CFI compares the fit of the hypothesized model to a baseline independence model in which all variables are uncorrelated. Values of CFI closer to 1.0 indicate a better fit, with values above 0.95 generally considered to reflect good model fit; values greater than .90 are thought to reflect adequate fit. The RMSEA evaluates the discrepancy between the hypothesized model and the population covariance matrix, adjusted for model complexity; values greater than .90 are thought to reflect adequate fit. RMSEA values below 0.05 are considered indicative of good fit, and values up to 0.08 are adequate. Lastly, SRMR was used to quantify the standardized difference between observed and predicted correlations, with lower values reflecting better fit,

with those values being less than 0.05 considered good and less than 0.08 as adequate. Chi-square values approximating their degrees of freedom and a non-significant p-value were also critical in confirming our model's adequacy. Modification indices were also reviewed to identify potential areas of model misfit and guide adjustments to improve the overall structure while maintaining theoretical constancy.

CFA Variants Depending on Assumptions. CFA has evolved significantly since its early formulation under the maximum likelihood (ML) estimation framework (Jöreskog, 1969) and now includes a diverse range of analytic approaches that account for non-normality, categorical data, and complex sample designs (Muthén & Satorra, 1995; Boomsma & Bollen, 1991). These advancements have expanded CFA's applicability beyond simple continuous-variable models, allowing researchers to better estimate model parameters and fit indices under real-world conditions such as ordinal scales, clustered data, or robust estimation needs.

The "basic" approach for CFA analyses is the Maximum Likelihood (ML) method, which assumes multivariate *normality* of *continuous* data and provides estimates that efficiently use all available information. However, the applicability of ML is limited when the data are ordinal or when normality assumptions are violated, which is the norm for Likert scale ratings common to measures designed for identification**.** Thus, while ML provides accurate estimates with normal and continuous, its performance can deteriorate with non-normal or non-continuous data. As such, numerous alternative estimation methods have been developed, including ML-centered adjustments, and non-ML centered adjustments (Li, 2016).

To address non-parametric data, a number of ML centered adjustments have been developed over time, including: scale shifted (MLs), robust (MLr), and categorical (MLc) estimations methods, each developed to address violations of the standard ML assumptions. The

scale-shifted method (MLs), which is often referred to as “means adjusted” approach, adjust the chi-square test statistics using the sample mean to account for non-normality in the data. The robust method (MLr), a “variance-adjusted” approach, goes a step further by correcting both the chi-square statistic and standard errors using the sample variance. This allows for the detection of potential outliers and better accounts for issues like skewed distributions and excess kurtosis (Savalei & Rhemtulla, 2013). These conditions are specifically meant to adjust for non-normality and improve the accuracy of model estimation and fit.

The categorical ML approach (MLc) was reflected in our use of polychoric correlations under the ML-ordered condition. This estimates the relationship between unobserved, continuous laten variable and underlie ordinal responses, rather than assuming linear relationships as in Pearson correlations. While we did not use a formal MLc estimator, our ML-ordered models incorporated polychoric correlations to better align with the ordinal data of the CSBQ. One thing to note is that we were unable to implement scaled shifted or robust correlations under this condition due to limitations in combining polychoric matrices with these adjustments. This presented a methodological challenge, as polychoric-based models assume a latent normal distribution that can be sensitive to data that are highly skewed or contain sparse categories (Long & Cheng, 2004).

Two of the most frequently recommended non-ML centered adjustments include categorical Diagonally Weighted Least Squares (DWLS) and Unweighted Least Squares (ULS), each of which uses different matrix weighting approaches (Salvalei & Rhemtulla, 2013). Central to DWLS and ULS is the polychoric correlation matrix, which estimated the relationship between unobserved continuous laten variables that lie beneath the ordinal responses. Both approaches assume that each ordinal item reflects an underlying normal continuum, which

allows the CFA model to more accurately estimate latent constructs from the Likert-type scale. DWLS is the most commonly used estimator in this class and applies a diagonal weight matrix based on the estimated variances of the polychoric correlations (Savalei & Rhemtulla, 2012). ULS minimizes the unweighted sum of squared differences between observed and model-implied correlations. This simplicity makes ULS particularly attractive when estimation becomes unstable, such as in small samples or when item thresholds are highly skewed. Although it's less commonly used than DWLS, ULS has shown slightly better performance in simulation studies, particularly in terms of convergence rates and the accuracy of parameter estimates (Savalei & Rhemtulla, 2012).

Utilizing Modification Indices to Change Models and Get Good Fit. Scale development is a multi-iterative process wherein numerous decisions are made (Manata & Boster, 2024). This leads to frequent engagement in questionable research practices (QRPS) including, among others non-transparent engagement with modification indices to make models "fit well" (Crede & Harms, 2019). Throughout the process, we examined modification indices to pinpoint potential model misfits, allowing for theoretical adjustments to enhance the model's structure(DiStefano & Morgan, 2014). This oftentimes can come across as quite minor adjustments to 1 or 2 modeled relationships between items and factors or items and items (i.e., correlating residuals). However, this practice has been described by leading methodologists as using a "rusty knife" to "fix" a measure that might mask foundational challenges to the theoretical conceptualization of the item-factor relationships. In the current study, we quantify the number of correlated residuals within hypothesized factors required to get poorly fitting model fit indices to fall into the "adequate range" (**Table 2**). Figure 1 provides an overall model of the different iterations of model testing followed by engaging in QRPs to "fix" items.

Table 2. Hypothesized and covarying residual models compared across different methods.

	Maximum Likelihood				Least Squares: Diagonally Weighted			Least Squares: Unweighted		
	ML	MLs	MLr	MLc	DWLS	DWLSs	DWLSr	ULS	ULSs	ULSr
Hypothesized Model (Hartman et al., 2006)	x	x	x	x	x	x	x	x	x	x
Adapted Model Allowing Covarying Items	x	x	x	x	x	x	x	x	x	x

Notes. ML = Maximum Likelihood, DWLS = Diagonally Weighted Least Squares, ULS= Unweighted Least Squares, Subscript $_s$ = Scale Shifted, Subscript $_r$ = Robust, Subscript $_c$ = Categorical.

In Step 1, each model began with a hypothesized factor structure where items were constrained to load on their intended factors, and all residuals were fixed to zero. No cross-loadings or residual covariances were included at this point. In Step 2, model fit was evaluated using established thresholds (CFI ≥ .90, RMSEA ≤ .06, SRMR ≤ .06). If model fit was inadequate, item residuals were allowed to covary within factors, not across, based on theoretical justification; stopping rules for this process included (a) modification indices (MI) guided the decision making until the model was either "fixed" (i.e., indices were all in the "good" range) or (b) MI reaching MI =100. In Step 3, we compared the original models to those with residual covariances added. Improvements in model fit were evaluated by observing increases in CFI and decreases in RMSEA and SRMR. Finally, in Step 4, we descriptively compared model performance across all estimation methods and residual conditions, summarizing global fit indices and the extent of residual correlations used. This allowed us to evaluate trade-offs between model parsimony and fit improvement, which are summarized in **Table 2**. Overall, the iterative process provided a structured and theory-driven approach to improving model performance without overfitting, and highlighted differences in how estimation methods respond to adjustments for item residuals.

Figure 1. Iterative steps for model testing and item residual "fixing."

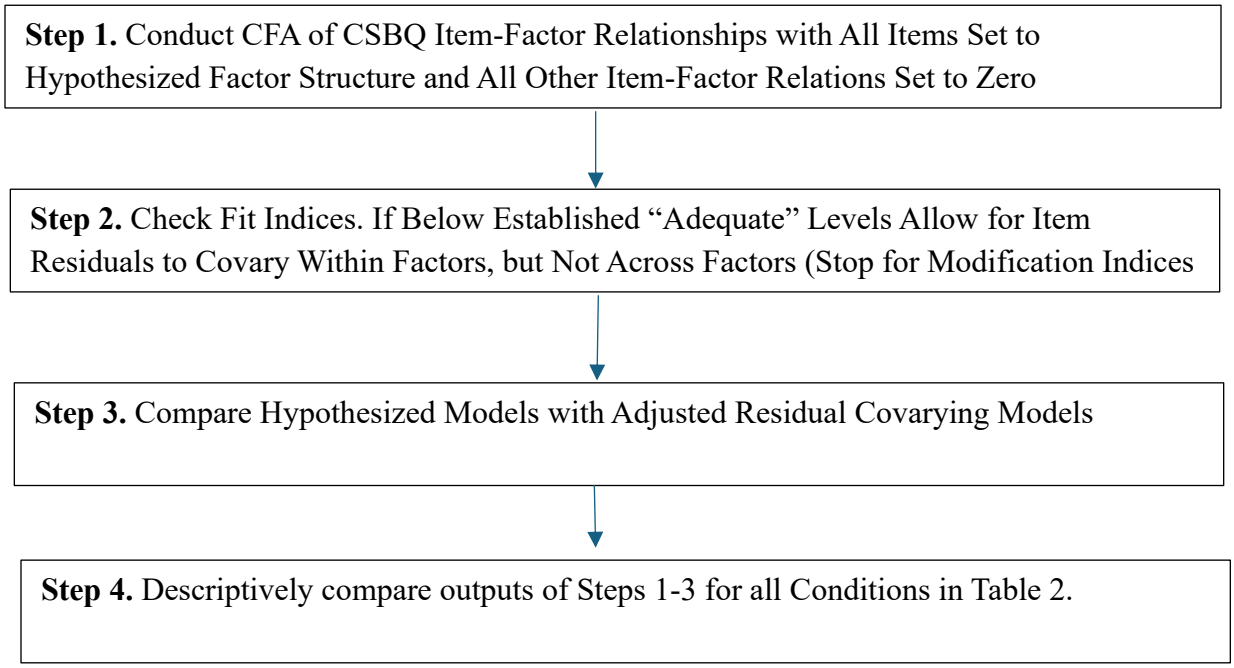

Figure 1 outlines the iterative process used to test and refine the confirmatory factor analysis.

Results

The sample consisted of 2,570 children aged 3 to 18 years (M = 8.9, SD = 3.47). Of these, 1,878 (73.0%) were male and 692 (27.0%) were female. Diagnostic subgroups included children with PDDNOS (n = 544), ADHD (n = 586), PDDNOS + ADHD (n = 164), High Functioning Autism (HFA; n = 102), Mental Retardation with PDD (MR+PDD; n = 152), Mental Retardation only (MR; n = 581), Internalizing Disorders (ID; n = 209), and Normal Control (NC; n = 232). A detailed breakdown of the diagnostic subgroups, including gender distribution and age characteristics, is provided in **Table 3.**

Table 3. Descriptive Characteristics of Diagnostic Subgroups in Hartman et al. (2006)

Diagnostic Group	**N**	**% Male**	**% Female**	**Age Range (Years)**	**Age Mean & (SD)**
PDDNOS	544	78%	22%	4–18	8.99 (3.47)
ADHD	586	86%	14%	4–18	8.84 (3.35)
PDDNOS + ADHD	164	87%	13%	4–16	7.85 (2.73)
High Functioning Autism (HFA)	102	85%	15%	4–18	9.42 (4.08)
Mental Retardation + PDD (MR+PDD)	152	76%	24%	3–18	11.22 (3.88)
Mental Retardation Only (MR)	581	59%	41%	4–18	11.98 (3.59)
Internalizing Disorders (ID)	209	49%	51%	4–18	10.90 (3.46)
Normal Control (NC)	232	45%	55%	4-18	8.34 (2.67)
Total	2570				

Unadjusted Models

Maximum Likelihood (ML). The standard ML model (M1) demonstrated poor to adequate/good fit, with a CFI of 0.830, a RMSEA of 0.058, and SRMS of 0.055 (**Table 4).** After allowing 21 item-pair residuals to correlate within factors (M1.1), model fit improved in the adequate to good range with a CFI of 0.900, RMSEA of 0.046, and SRMS of 0.048.

Maximum Likelihood-Categorical. The ML-ordered model (M2.0), estimated using polychoric correlations for ordinal data, demonstrated poor model fit, with a CFI of 0.771,

RMSEA of 0.086, and SRMR of 0.067 (**Table 4**). After allowing residuals among 61 item pairs (M2.1), model fit improved substantially to adequate to good, with a CFI of 0.915, RMSEA of 0.042, and SRMR of 0.046.

Diagonally Weighted Least Squares (DWLS). The DWLS standard approach (M3.0) demonstrated a good to excellent fit with a CFI of 0.977, RMSEA of 0.063, and SRMS of 0.061 (**Table 4**). After allowing residual correlations among 16 item pairs within factors (M3.2), model fit improved further into the good to adequate range, yielding a CFI of 0.983, RMSEA of 0.054, and SRMR of 0.055.

Unweighted Least Squares (ULS).- The ULS model (M4.0) demonstrated good to excellent fit, with a CFI of 0.976, RMSEA of 0.060, and SRMR of 0.060 (**Table 4**). Allowing residual correlations among 9 item pairs (M4.2) resulted in slightly improved model fit in the poor to adequate range, with a CFI of 0.887, RMSEA of 0.051, and SRMR of 0.056.

Scale Shifted

Maximum Likelihood (ML)- The scale-shifted ML model (M1) demonstrated slightly improved fit over the unadjusted scaled ML model, with a CFI of 0.828, RMSEA of 0.053, and SRMR of 0.055 (**Table 4**). After allowing 21 item-pair residuals to correlate within factors (M1.2), model fit improved to the adequate to good range, with a CFI of 0.896, RMSEA of 0.042, and SRMR of 0.048. which can suggest that adjusting the chi-square statistics to account for non-normality helped improve model fit.

Maximum Likelihood-Categorical.- Scale-shifted estimation could not be conducted under the ML-ordered condition due to incompatibility between polychoric correlation matrices and mean-adjusted test statistics (**Table 4**). This limitation arises because polychoric estimation

assumes latent normality, which conflicts with the variance-correcting adjustments required by scale-shifted methods.

Diagonally Weighted Least Squares (DWLS).- The scale-shifted DWLS model (M3.1) demonstrated a decline in fit compared to the unadjusted model, with a CFI of 0.877, RMSEA of 0.062, and SRMR of 0.061 (**Table 4**). After allowing residual correlations among 8 item pairs (M3.2), model fit improved to adequate to good range, with a CFI of 0.908, RMSEA of 0.054, and SRMR of 0.055. To further assess model sensitivity, we ran a follow-up model (M3.3) allowing correlations among all 16 item pairs initially flagged, which brought both RMSEA and SRMR below the pre-specified cutoff values. However, this change also resulted in a slight decrease in CFI from 0.908 to 0.900, and a slight increase in RMSEA (0.056) and SRMR (0.057). These shifts reflect a modest trade-off between improved residual fit and diminished comparative model fit, suggesting diminishing returns in model complexity when allowing additional residual correlations.

Unweighted Least Squares (ULS).- The scale-shifted ULS model (M4.1) demonstrated a modest decrease in fit, with a CFI of 0.872, RMSEA of 0.054, and SRMR of 0.060 (**Table 4**). After allowing 9 item-pair residuals to correlate (M4.2), fit improved slightly to the poor to adequate range with a CFI of 0.887, RMSEA of 0.051, and SRMR of 0.056, indicating that while the model remained robust, scaling adjustments introduced some instability.

Robust

Maximum Likelihood (ML)- The robust ML model (M1) showed marginal improvement in CFI and comparable RMSEA relative to the unadjusted version, with a CFI of 0.833, RMSEA

of 0.058, and SRMR of 0.055 (**Table 4**). After allowing 21 item-pair residuals to correlate (M1.2), model fit improved to adequate to good range, with a CFI of 0.902, RMSEA of 0.045, and SRMR of 0.048.

Maximum Likelihood – Categorical (ML-Ordered)-Robust estimation was not conducted under the ML-ordered condition due to the same restrictions as the scale-shifted model—namely, the incompatibility between polychoric correlation matrices and robust test statistic estimation (**Table 4**).

Diagonally Weighted Least Squares (DWLS)- The robust DWLS model (M3.1) demonstrated substantial deterioration in fit, with a CFI of 0.784, RMSEA of 0.082, and SRMR of 0.061 (**Table 4).** After allowing 16 item-pair residuals to correlate (M3.2), model fit improved moderately to the adequate to poor range CFI of 0.853, RMSEA of 0.068, and SRMR of 0.055.

Unweighted Least Squares (ULS)- The robust ULS model (M4.1) exhibited poor fit, with a CFI of 0.782, RMSEA of 0.082, and SRMR of 0.060 (**Table 4**). After allowing 9 residual correlations (M4.2), the fit improved slightly to the adequate to poor range with a CFI of 0.829, RMSEA of 0.073, and SRMR of 0.056, though still indicating difficulties in managing robust adjustments within this estimation approach.

Table 4. Factor Analysis for ML, ML-Ordered, DWLS and ULS Results in Hartman et al. (2006)

Unadjusted		CFI	RMSEA	SRMS
Maximum Likelihood (ML)	**M1**	**.830**	**.058**	**.055**
ML – Correlated Items	**M1.1**	**.898**	**.046**	**.048**
ML - Ordered	**M2.0**	**.771**	**.086**	**.067**

ML- Ordered Correlated Items	**M2.1**	**.915**	**.042**	**.046**
Diagonally Weighted Least Squares (DWLS)	**M3.0**	.977	.063	.061
DWLS- Correlated Items	**M3.2**	.981	.056	.057
Unweighted Least Squares (ULS)	**M4.0**	.976	.060	.060
ULS- Correlated Items	**M4.2**	.979	.056	.056
Scale and Scale Shifted		Scaled and Shifted CFI	Scaled and Shifted RMSEA	SRMS
Maximum Likelihood (ML)	**M1**	**.828**	**.053**	**.055**
ML – Correlated Items	**M1.2**	**.896**	**.042**	**.048**
Diagonally Weighted Least Squares (DWLS)	**M3.1**	.877	.062	.061
DWLS- Correlated Items	**M3.2**	.908	.054	.055
8 out 16 correlated	**M3.3**	.900	.056	.057
Unweighted Least Squares (ULS)	**M4.1**	.872	.054	.060
ULS- Correlated Items	**M4.2**	.887	.051	.056
Robust		Robust CFI	Robust RMSEA	SRMS
Maximum Likelihood (ML)	**M1**	**.833**	**.058**	**.055**
ML – Correlated Items	**M1.2**	**.902**	**.045**	**.048**
Diagonally Weighted Least Squares (DWLS)	**M3.1**	.784	.082	.061
DWLS- Correlated Items	**M3.2**	.853	.068	.055
Unweighted Least Squares (ULS)	**M4.1**	.782	.082	.060
ULS- Correlated Items	**M4.2**	.829	.073	.056

Note: CFI- Comparative Fit Index, RMSEA-Root Mean Square Error of Approximation, SRMS-Standardized Root Mean Square Residual

Discussion

This study evaluated the structural validity of the Children's Social Behavior Questionnaire (CSBQ) using confirmatory factor analysis (CFA) across multiple estimation

approaches. Specifically, we compared traditional Maximum Likelihood (ML), ML with polychoric correlations (ML-Ordered), Diagonally Weighted Least Squares (DWLS), and Unweighted Least Squares (ULS) models under unadjusted, scale-shifted, and robust conditions. Our goal was to explore how different estimation methods and data assumptions affect model fit, especially given the ordinal nature of the CSBQ and the potential for non-normality in clinical screening tools that may be used as outcome variables in a manner for which they are not designed. Initial unadjusted models revealed considerable variation in performance depending on the estimation method used.

Model Fit Across Estimators

The standard Maximum Likelihood (ML) model demonstrated poor-to-adequate fit, while ML-Ordered estimation yielded even poorer initial results. This is likely attributable to the ordinal structure of the CSBQ, where ML fails to accommodate the categorical nature of the data and the presence of non-normality. Mîndrilă (2010) found that ML performs well only when variables are continuous and normally distributed, and its use with ordinal or non-normally distributed data leads to inflated chi-square values and biased fit indices such as RMSEA and CFI. These shortcomings were particularly pronounced in our sample, where the ML-Ordered model initially failed to capture the structural complexity of the CSBQ until residual covariances were added. Once correlated residuals were introduced, the ML-Ordered condition showed the most dramatic improvement in fit (CFI increasing from .771 to .915), underscoring the presence of localized item-level misfit.

In contrast to ML, both DWLS and ULS estimation produced good-to-excellent model fit in their initial unadjusted forms, aligning with findings from Mîndrilă (2010) that DWLS provides more accurate parameter estimates and robust model fit in the presence of ordinal and

non-normal data. Unlike ML, DWLS is based on polychoric correlations and uses the asymptotic variance to correct for standard errors and model fit statistics, which enhances its applicability to Likert-type scales commonly used in ASD research. When robust and scale-shifted corrections were applied, DWLS and ULS remained more stable than ML; however, these corrections resulted in decreases in CFI and increases in RMSEA, particularly for DWLS. Considering that scale-shifted and robust approaches are considered more optimal, these findings suggest that these adjustments reveal poor item to factor relationships for the CSBQ.

DWLS deserves particular emphasis in this context. It is widely regarded as best practice in autism-related scale validation due to its suitability for modeling ordinal responses, small sample sizes, and skewed distributions—conditions that frequently occur in developmental and clinical datasets. In our study, applying all 16 flagged correlated residuals in the DWLS scale-shifted condition resulted in only marginal improvement in RMSEA and SRMR, with a slight decrease in CFI (from 0.908 to 0.900). This pattern reflects diminishing returns from increasing model complexity and highlights the importance of maintaining parsimony. As Mîndrilă (2010) cautioned, often DWLS estimates require numerous modification indices that, if addressed via engagement in the QRP of correlating residuals, may obscure the true item-factor latent structure.

These findings have three key implications. First, for CSBQ development, these findings suggest that DWLS is a methodologically sound approach for confirming factor structures in tools designed to measure subtle social and behavioral traits in autistic children. However, the fact that the CSBQ achieves a strong model fit with limited adjustments with the unadjusted approach but poor fit with the more appropriate DWL scale-shifted approach displays the ease with which poorly performing scales can be *presented as adequate* when in fact more work is required in scale development. Second, using ML-based estimators without addressing their

limitations could result in over- or underestimation of intervention effects. Finally, for ASD identification tools, these results suggest that early screening and diagnostic instruments that rely on ordinal scales—such as the CSBQ or others like the ADOS or ADI-R—should be routinely validated using scale shifted DWLS approaches to ensure their measurement properties are reliable, especially when arguing they should be used as outcome variables.

Significance

The significance of this study lies in its implications for how autism-related behavioral assessments, such as the Children's Social Behavior Questionnaire (CSBQ), are interpreted and applied in public health, clinical, and research contexts. Autism spectrum disorder (ASD) remains a rapidly evolving diagnostic area, with increasing emphasis on early detection and personalized intervention (Chahin et al., 2020). As screening and diagnostic tools are increasingly used in pediatric primary care and large-scale public health surveillance, the statistical integrity of the tools used becomes a public health concern. Our results confirm that confirmatory factor analysis (CFA) model outcomes are sensitive not only to the item content of the tool, but also to the choice of estimation method—particularly when applied to ordinal data common in ASD screeners and outcome measures.

From a methodological perspective, this study supports a shift away from traditional Maximum Likelihood (ML) estimation methods, which assume continuous and normally distributed data. These assumptions are often violated in autism screening instruments that rely on Likert-type ratings of behaviors. As Mîndrilă (2010) demonstrated, ML estimation produces inflated chi-square statistics and biased standard errors when used with ordinal and non-normal data, which can ultimately distort model fit and obscure valid constructs. Our study corroborates this concern by showing poor-to-adequate model performance using ML and ML-Ordered

estimation techniques, with significant reliance on residual correlations to improve fit. In contrast, estimation methods tailored to ordinal structures—such as Diagonally Weighted Least Squares (DWLS) and Unweighted Least Squares (ULS)—produced superior initial fit and required fewer modifications, but poor fit when using the more appropriate scale shifted approaches. This reinforces the need for a critical evaluation of methodological justification for common psychometric tools designed as predictors but analyzed as outcomes in ASD research.

Clinically, these findings are critical for professionals using tools like CSBQ not just for initial identification, but also to monitor behavioral changes following interventions. Chahin et al. (2020) underscore that early and repeated screening is now routine in pediatric settings, where screeners like the CSBQ are used to determine access to interventions such as Applied Behavior Analysis (ABA), occupational therapy, or speech-language services. If the factor structure of these screeners is misrepresented due to inappropriate analytic approaches, this can lead to inaccurate conclusions about a child's developmental progress or the effectiveness of treatment. This is particularly problematic given the push for universal screening between 18 and 24 months and the known variability in ASD symptom profiles and comorbidities (Chahin et al., 2020). Ultimately, tools that have good evidence as predictors of outcomes in diagnostic accuracy studies may ultimately have minimal evidence supporting their use as outcome variables. Unless evidence suggests that tools are clearly designed as psychometrically valid outcome variables, interventions using clinical scales designed for identification may be using screening/assessment tools adequate for identification but inappropriate as outcomes.

From a policy standpoint, reliance on valid and replicable statistical findings is essential for funding decisions, resource allocation, and public reporting of health outcomes. Screening tools that are improperly used as outcomes may fail to capture nuanced behavioral traits that

might inform personalized care, or be thought to measure univariate outcomes when in fact they are confounded by multivariate influences. The present study emphasizes that statistical misalignment—whether the use of ML methods with ordinal data or DWLS without mean adjustments—could lead to misrepresentation of the item-factor relationships when, in fact, more research is needed regarding the adequacy of a tool as an outcome variable.

In summary, the results highlight that analytic decisions—particularly the use of ordinal-specific estimators like scale shifted DWLS—are more than statistical "preferences;" they are critical to the valid interpretation and ethical application of autism screening instruments. This aligns with recommendations from both Chahin et al. (2020) and Mîndrilă (2010), calling for increased methodological rigor in how psychometric tools are used to inform both individual care and population-level decision-making.

Implications and Methodological Considerations

Our results reinforce prior findings that Diagonally Weighted Least Squares (DWLS) and Unweighted Least Squares (ULS) are more appropriate for analyzing ordinal data (Savalei & Rhemtulla, 2012). These estimation techniques, unlike traditional Maximum Likelihood (ML), are specifically designed to handle the non-normal, categorical nature of Likert-type data, which is commonly found in clinical and developmental assessments such as the Children's Social Behavior Questionnaire (CSBQ). In this study, ML-based estimators—particularly the ML-Ordered model—were more prone to misfit, requiring multiple post-hoc residual correlations to meet acceptable model thresholds. This is consistent with findings from Mîndrilă (2010), who demonstrated that ML estimation applied to ordinal or non-normally distributed variables results in biased chi-square values, underestimated standard errors, and poor fit indices, leading to distorted conclusions about the underlying factor structure.

Importantly, this study also revealed a methodological constraint: the inability to apply scale-shifted or robust corrections to ML-Ordered models due to current software limitations. Although polychoric correlations are more appropriate for ordinal indicators, they are not compatible with commonly used robust estimation procedures such as the Satorra-Bentler correction (Muthén & Satorra, 1995). This creates a critical trade-off—researchers must often choose between using estimators aligned with their data structure or applying corrections that account for distributional violations. The lack of flexibility in this regard limits the robustness and reproducibility of CFA models built on ordinal data, particularly in high-stakes applied research settings.

These methodological implications are especially relevant as autism screening tools are increasingly integrated into pediatric care, school evaluations, and public health monitoring systems as outcomes of interest. Chahin et al. (2020) highlights the importance of repeated, early screening and the role that validated assessment instruments play in ensuring timely and equitable intervention. If the psychometric properties of these instruments are misrepresented due to misaligned analytic approaches, there is a risk of using scales in a manner for which they are not designed.

To strengthen the future of measurement modeling in this area, statistical software must evolve to support robust estimation procedures that can be combined with polychoric inputs. This would remove a key barrier and enable researchers to accurately model ordinal data while addressing non-normality and localized misfit. Additionally, newer estimation approaches such as Bayesian modeling or item response theory (IRT) may offer further flexibility and insight for tools with complex structures or multidimensionality. Eventually, the adoption of estimation strategies that are both methodologically and philosophically aligned with the nature of the data

is essential for preserving the validity, fairness, and real-world relevance of autism screening and behavioral assessment tools.

Limitations and Future Directions

Several limitations should be acknowledged in interpreting the results of this study. First, while polychoric correlations are widely used and recommended for modeling ordinal data in confirmatory factor analysis, they rely on the assumption that observed ordinal variables reflect underlying continuous and bivariate normally distributed traits. This assumption may not hold uniformly across all CSBQ items, especially given the variability in behavioral expression in children with autism spectrum disorder. Violations of this assumption likely contributed to the convergence difficulties observed in the ML-Ordered models when scale-shifted and robust corrections were attempted. Second, although the inclusion of residual covariances improved model fit across multiple estimation conditions, these modifications may reflect sample-specific noise rather than theoretically meaningful relationships. Overreliance on modification indices to improve fit risks reducing construct clarity and replicability. Future studies should use cross-validation or multi-sample CFA to determine whether residual correlations represent generalizable aspects of the measure's structure or are idiosyncratic to a particular dataset. Third, this study was constrained by software limitations that prevented the implementation of scale-shifted or robust corrections for models using polychoric correlations. This technical limitation restricts researchers' ability to evaluate model robustness when using estimation techniques that are otherwise well-aligned with the ordinal nature of psychometric data. As noted by Mîndrilă (2010), such constraints can result in a methodological compromise where researchers must choose between modeling data correctly and adjusting for non-normality. Until statistical

software packages integrate these capabilities more fully, a complete understanding of model performance under ideal conditions remains difficult.

Looking ahead, future research should explore several promising directions. First, studies should replicate these findings using independent samples and consider multi-group CFA to test the stability of factor structures across demographic subgroups such sex, race/ethnicity, and diagnostic status. Second, alternative estimation frameworks such as Bayesian CFA or item response theory (IRT) models may offer greater flexibility in modeling complex response structures, handling missing data, and quantifying parameter uncertainty. Third, there is a need for continued refinement of autism-specific measurement tools like the CSBQ to ensure that they are psychometrically sound across developmental stages and diverse populations. These advancements are essential not only for academic rigor but also for supporting valid, fair, and generalizable interpretations in clinical and policy-relevant contexts. Finally, there is a need to better understand the interplay between adequate item-factor analyses for outcome variables and diagnostic accuracy. Currently, these analyses only speak to challenges related to the common practice of using tools primarily designed for diagnostic identification purposes as outcome variables.

Conclusion

This study examined how different confirmatory factor analysis (CFA) estimation methods—Maximum Likelihood (ML), ML for ordinal data (ML-Ordered), Diagonally Weighted Least Squares (DWLS), and Unweighted Least Squares (ULS)—impact model fit and interpretability of the Children's Social Behavior Questionnaire (CSBQ), a widely used tool in autism screening and outcome research. Our results showed that estimation methods aligned with the ordinal nature of the data, particularly DWLS and ULS, consistently produced superior model fits and required fewer post-hoc modifications than ML-based approaches. In contrast, ML and ML-Ordered models were more sensitive to misfit, requiring residual correlations and performing poorly under robust correction conditions due to both methodological and software limitations.

These findings highlight a critical methodological insight: analytic decisions—especially the selection of estimation techniques—directly shape the interpretation of psychometric tools and, by extension, influence the clinical, research, and policy decisions derived from them. DWLS and ULS, by considering the ordinal structure of the CSBQ, offered more stable and interpretable models, supporting their use in autism-related scale development and validation efforts. Importantly, this study reinforces the need for researchers, clinicians, and public health professionals to adopt estimation techniques that are both statistically rigorous and philosophically aligned with the purpose of the scales. Failure to do so may not only compromise model validity but also risk introducing bias into how behavioral traits are measured, interpreted, and acted upon. As autism screening tools are increasingly embedded in early intervention frameworks as outcomes, the quality of the psychometric evidence base supporting the use of these instruments as outcomes is essential for ensuring equitable and evidence-based care.

References

Akshoomoff, N., Corsello, C., & Schmidt, H. (2006). The Role of the Autism Diagnostic Observation Schedule in the Assessment of Autism Spectrum Disorders in School and Community Settings. *The California School Psychologist*, *11*(1), 7–19. https://doi.org/10.1007/BF03341111

Alpert, J. S. (2021). Autism: A Spectrum Disorder. *The American Journal of Medicine*, *134*(6), 701–702. https://doi.org/10.1016/j.amjmed.2020.10.022

American Academy of Pediatrics. (2006). Identifying infants and young children with developmental disorders in the medical home: An algorithm for developmental surveillance and screening. *Pediatrics*, *118*(1), 405–420

American Psychological Association (2013). *Autism Spectrum Disorder*. In Diagnostic and Statistical Manual of Mental Disorders (5th ed.). https://www.apa.org/topics/autism-spectrum-disorder

Barger, B. (2024). Epidemiology with psychometric spirit: MoBa leads autism's interdisciplinary future—A commentary on Havdahl et al. (2023). *Journal of Child Psychology and Psychiatry, 65*(8), 1115–1118. https://doi.org/10.1111/jcpp.13933

Barger, B., Rice, C., Simmons, C. A., & Wolf, R. (2018). A Systematic Review of Part C Early Identification Studies. *Topics in early childhood special education*, *38*(1), 4–16. https://doi.org/10.1177/0271121416678664

Barger, B., Salmon, A., & Moore, Q. (2024). Medical Home, Developmental Monitoring/Screening, and Early Autism Identification. *Journal of autism and developmental disorders*, *54*(9), 3242–3261. https://doi.org/10.1007/s10803-023-06044-0

Bishop, S.L. and Lord, C. (2023), Commentary: Best practices and *processes* for assessment of autism spectrum disorder – the intended role of standardized diagnostic instruments. J Child Psychol Psychiatr, 64: 834-838. https://doi.org/10.1111/jcpp.13802

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford publications.

Carbone, P. S., Campbell, K., Wilkes, J., Stoddard, G. J., Huynh, K., Young, P. C., & Gabrielsen, T. P. (2020). Primary Care Autism Screening and Later Autism Diagnosis. *Pediatrics*, *146*(2), e20192314. https://doi.org/10.1542/peds.2019-2314

Casson, R. J., & Farmer, L. D. (2014). Understanding and checking the assumptions of linear regression: a primer for medical researchers. *Clinical & experimental ophthalmology*, *42*(6), 590–596. https://doi.org/10.1111/ceo.12358

Chahin, S. S., Apple, R. W., Kuo, K. H., & Dickson, C. A. (2020). Autism spectrum disorder: psychological and functional assessment, and behavioral treatment approaches. *Translational pediatrics*, *9*(Suppl 1), S66–S75. https://doi.org/10.21037/tp.2019.11.06

Chesnut, S. R., Wei, T., Barnard-Brak, L., & Richman, D. M. (2017). A meta-analysis of the social communication questionnaire: Screening for autism spectrum disorder. *Autism : the international journal of research and practice*, *21*(8), 920–928. https://doi.org/10.1177/1362361316660065

Clements, C. C., Watkins, M. W., Schultz, R. T., & Yerys, B. E. (2020). Does the Factor Structure of IQ Differ Between the Differential Ability Scales (DAS-II) Normative Sample and Autistic Children?. *Autism research : official journal of the International Society for Autism Research*, *13*(7), 1184–1194. https://doi.org/10.1002/aur.2285

Constantino, J. N., Gruber, C. P., Davis, S., Hayes, S., Passanante, N., & Przybeck, T. (2004). The factor structure of autistic traits. *Journal of child psychology and psychiatry, and allied disciplines*, *45*(4), 719–726. https://doi.org/10.1111/j.1469-7610.2004.00266.x

Crede, M., & Harms, P. (2019). Questionable research practices when using confirmatory factor analysis. *Journal of Managerial Psychology, 34*(1), 18–30. https://doi.org/10.1108/JMP-06-2018-0272

Dash, G., & Paul, J. (2021). CB-SEM vs PLS-SEM methods for research in social sciences and technology forecasting. *Technological Forecasting and Social Change*, *173*, 121092. https://doi.org/10.1016/j.techfore.2021.121092

de Bildt, A., Mulder, E. J., Hoekstra, P. J., Van Lang, N. D. J., Minderaa, R. B., & Hartman, C. A. (2009). Validity of the Children's Social Behavior Questionnaire (CSBQ) in Children with Intellectual Disability: Comparing the CSBQ with ADI-R, ADOS, and Clinical DSM-IV-TR Classification. *Journal of Autism and Developmental Disorders*, *39*(10), 1464–1470. https://doi.org/10.1007/s10803-009-0764-x

de Bildt, A., Sytema, S., Ketelaars, C., Kraijer, D., Volkmar, F., & Minderaa, R. (2003). Measuring Pervasive Developmental Disorders in Children and Adolescents with Mental Retardation: A Comparison of Two Screening Instruments Used in a Study of the Total Mentally Retarded Population from a Designated Area. *Journal of Autism and Developmental Disorders*, *33*(6), 595–605. https://doi.org/10.1023/B:JADD.0000005997.92287.a3

De La Roche, L., Derby, B., Pascoe, M., Susko, M., Lutchmeah, S., Jones, J., Georgiades, S., Nicolson, R., Anagnostou, E., & Kelley, E. (2024). Brief report: A confirmatory factor

analysis of the Child Behavior Checklist in a large sample of autistic youth. *Research in Autism Spectrum Disorders*, *118*, 102487. https://doi.org/10.1016/j.rasd.2024.102487

DiStefano, C., & Morgan, G. B. (2014). A Comparison of Diagonal Weighted Least Squares Robust Estimation Techniques for Ordinal Data. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(3), 425–438. https://doi.org/10.1080/10705511.2014.915373

DeVellis, R.F. and Thorpe, C.T. (2021) Scale Development: Theory and Applications. Sage Publications, Thousand Oaks.

Distinguishing between screening and assessment for mental and behavioral health problems. (2014) https://www.apaservices.org/practice/reimbursement/billing/assessment-screening

Dumont-Mathieu, T., & Fein, D. (2005). Screening for autism in young children: The Modified Checklist for Autism in Toddlers (M-CHAT) and other measures. *Mental retardation and developmental disabilities research reviews*, *11*(3), 253–262. https://doi.org/10.1002/mrdd.20072

Falkmer, T., Anderson, K., Falkmer, M., & Horlin, C. (2013). Diagnostic procedures in autism spectrum disorders: a systematic literature review. *European child & adolescent psychiatry*, *22*(6), 329–340. https://doi.org/10.1007/s00787-013-0375-0

Farmer, C., Kaat, A. J., Edwards, M. C., & Lecavalier, L. (2024). Measurement Invariance in Intellectual and Developmental Disability Research. *American journal on intellectual and developmental disabilities*, *129*(3), 191-198. https://doi.org/10.1352/1944-7558-129.3.191

Frigaux, A., Evrard, R., & Lighezzolo-Alnot, J. (2019). L'ADI-R et l'ADOS face au diagnostic différentiel des troubles du spectre autistique : intérêts, limites et ouvertures [ADI-R and ADOS and the differential diagnosis of autism spectrum disorders: Interests, limits, and openings]. *L'Encephale*, *45*(5), 441–448. https://doi.org/10.1016/j.encep.2019.07.002

Grove, R., Begeer, S., Scheeren, A. M., Weiland, R. F., & Hoekstra, R. A. (2021). Evaluating the latent structure of the non-social domain of autism in autistic adults. *Molecular Autism*, *12*(1), 22. https://doi.org/10.1186/s13229-020-00401-x

Harris, R. P., Helfand, M., Woolf, S. H., Lohr, K. N., Mulrow, C. D., Teutsch, S. M., Atkins, D., & Methods Work Group, Third US Preventive Services Task Force (2001). Current methods of the US Preventive Services Task Force: a review of the process. *American journal of preventive medicine*, *20*(3 Suppl), 21–35. https://doi.org/10.1016/s0749-3797(01)00261-6

Hartman, C. A., Luteijn, E., Serra, M., & Minderaa, R. (2006). Refinement of the Children's Social Behavior Questionnaire (CSBQ): an instrument that describes the diverse problems seen in milder forms of PDD. *Journal of autism and developmental disorders*, *36*(3), 325–342. https://doi.org/10.1007/s10803-005-0072-z

Hiruma, L., Pretzel, R. E., Tapia, A. L., Bodfish, J. W., Bradley, C., Wiggins, L., Hsu, M., Lee, L. C., Levy, S. E., & Daniels, J. (2021). A Distinct Three-Factor Structure of Restricted and Repetitive Behaviors in an Epidemiologically Sound Sample of Preschool-Age Children with Autism Spectrum Disorder. *Journal of autism and developmental disorders*, *51*(10), 3456–3468. https://doi.org/10.1007/s10803-020-04776-x

Hobart, J. C., Cano, S. J., Zajicek, J. P., & Thompson, A. J. (2007). Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. *The Lancet. Neurology*, *6*(12), 1094–1105. https://doi.org/10.1016/S1474-4422(07)70290-9

Huerta, M., Bishop, S. L., Duncan, A., Hus, V., & Lord, C. (2012). Application of DSM-5 criteria for autism spectrum disorder to three samples of children with DSM-IV diagnoses of pervasive developmental disorders. *The American journal of psychiatry*, *169*(10), 1056–1064. https://doi.org/10.1176/appi.ajp.2012.12020276

Ikeda, E., Hinckson, E., & Krägeloh, C. (2014). Assessment of quality of life in children and youth with autism spectrum disorder: a critical review. *Quality of life research : an international journal of quality of life aspects of treatment, care, and rehabilitation*, *23*(4), 1069–1085. https://doi.org/10.1007/s11136-013-0591-6

Jones, W., & Klin, A. (2009). Heterogeneity and homogeneity across the autism spectrum: the role of development. *Journal of the American Academy of Child and Adolescent Psychiatry*, *48*(5), 471–473. https://doi.org/10.1097/CHI.0b013e31819f6c0d

Lebersfeld, J. B., Swanson, M., Clesi, C. D., & O'Kelley, S. E. (2021). Systematic Review and Meta-Analysis of the Clinical Utility of the ADOS-2 and the ADI-R in Diagnosing Autism Spectrum Disorders in Children. *Journal of autism and developmental disorders*, *51*(11), 4101–4114. https://doi.org/10.1007/s10803-020-04839-z

Li C. H. (2016). The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. *Psychological methods*, *21*(3), 369–387. https://doi.org/10.1037/met0000093

Long, J. S., & Cheng, S. (2004). Regression models for categorical outcomes. *Handbook of data analysis*, 259-284.

Lord, C., Rutter, M., & Le Couteur, A. (1994). Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of autism and developmental disorders*, *24*(5), 659–685. https://doi.org/10.1007/BF02172145

Lord, C., & Western Psychological Services. (2012). *ADOS-2 : autism diagnostic observation schedule : manual* (2nd ed.). Western Psychological Services.

Lord, C., Storoschuk, S., Rutter, M., & Pickles, A. (1993). Using the ADI-R to diagnose autism in preschool children. *Infant Mental Health Journal*, *14*(3), 234–252. https://doi.org/10.1002/1097-0355(199323)14:3<234::AID-IMHJ2280140308>3.0.CO;2-F

Luteijn, E., Luteijn, F., Jackson, S., Volkmar, F., & Minderaa, R. (2000). The children's Social Behavior Questionnaire for milder variants of PDD problems: evaluation of the psychometric characteristics. *Journal of autism and developmental disorders*, *30*(4), 317–330. https://doi.org/10.1023/a:1005527300247

Maenner, M. J., Warren, Z., Williams, A. R., Amoakohene, E., Bakian, A. V., Bilder, D. A., Durkin, M. S., Fitzgerald, R. T., Furnier, S. M., Hughes, M. M., Ladd-Acosta, C. M., McArthur, D., Pas, E. T., Salinas, A., Vehorn, A., Williams, S., Esler, A., Grzybowski, A., Hall-Lande, J., Nguyen, R. H. N., … Shaw, K. A. (2023). Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2020. *Morbidity*

and mortality weekly report. Surveillance summaries (Washington, D.C. : 2002), *72*(2), 1–14. https://doi.org/10.15585/mmwr.ss7202a1

Manata, B., & Boster, F. J. (2024). Reconsidering the Problem of Common-Method Variance in Organizational Communication Research. Management Communication Quarterly, 38(3), 677-697. https://doi.org/10.1177/08933189231226242

Marlow, M., Servili, C., & Tomlinson, M. (2019). A review of screening tools for the identification of autism spectrum disorders and developmental delay in infants and young children: recommendations for use in low- and middle-income countries. *Autism research : official journal of the International Society for Autism Research*, *12*(2), 176–199. https://doi.org/10.1002/aur.2033

Marsh, H. W., Guo, J., Dicke, T., Parker, P. D., & Craven, R. G. (2020). Confirmatory Factor Analysis (CFA), Exploratory Structural Equation Modeling (ESEM), and Set-ESEM: Optimal Balance Between Goodness of Fit and Parsimony. *Multivariate behavioral research*, *55*(1), 102–119. https://doi.org/10.1080/00273171.2019.1602503

Matsunaga, M. (2010). How to Factor-Analyze Your Data Right: Do's, Don'ts, and How-To's. *International journal of psychological research*, *3*(1), 97-110.

Mîndrilă, D. (2010). *Maximum Likelihood (ML) and Diagonally Weighted Least Squares (DWLS) estimation procedures: A comparison of estimation bias with ordinal and multivariate non-normal data*. *International Journal of Digital Society (IJDS)*, 1(1), 60–66. https://doi.org/10.20533/ijds.2040.2570.2010.0010

Moore, D. M., Baggett, K. M., & Barger, B. (2021). Measuring parent positive support of social communication among toddlers with autism: A systematic review. *Psychosocial Intervention, 30*(1), 57–66. https://doi.org/10.5093/pi2020a20

Muthen, B. O., & Satorra, A. (1995). Complex Sample Data in Structural Equation Modeling. *Sociological Methodology*, *25*, 267. https://doi.org/10.2307/271070

Newsom, J. (2008). *Some Clarifications and Recommendations on Fit Indices*. Psy 523/623 Structural Equation Modeling, Spring 2023. https://web.pdx.edu/~newsomj/semclass/ho_fit.pdf

Norris, M., & Lecavalier, L. (2010). Screening accuracy of Level 2 autism spectrum disorder rating scales. A review of selected instruments. *Autism : the international journal of research and practice*, *14*(4), 263–284. https://doi.org/10.1177/1362361309348071

Pfeiffer, S. I., & Blei, S. (2008). Gifted identification beyond the IQ test: Rating scales and other assessment procedures. In S. I. Pfeiffer (Ed.), *Handbook of giftedness in children: Psychoeducational theory, research, and best practices* (pp. 177–198). Springer Science + Business Media. https://doi.org/10.1007/978-0-387-74401-8_10

Randall, M., Egberts, K. J., Samtani, A., Scholten, R. J., Hooft, L., Livingstone, N., Sterling-Levis, K., Woolfenden, S., & Williams, K. (2018). Diagnostic tests for autism spectrum disorder (ASD) in preschool children. *The Cochrane database of systematic reviews*, *7*(7), CD009044. https://doi.org/10.1002/14651858.CD009044.pub2

Revelle, W. (2024). The seductive beauty of latent variable models: Or why I don't believe in the Easter Bunny. *Personality and Individual Differences, 221,* 1-17. https://doi.org/10.1016/j.paid.2024.112552

Robins, D. L., & Dumont-Mathieu, T. M. (2006). Early screening for autism spectrum disorders: update on the modified checklist for autism in toddlers and other measures. *Journal of developmental and behavioral pediatrics : JDBP*, *27*(2 Suppl), S111–S119. https://doi.org/10.1097/00004703-200604002-00009

Savalei V & Rhemtulla, M. (2013). The performance of robust test statistics with categorical data. *British Journal of Mathematical and Statistical Psychology, 66*(2), 201–223. https://doi.org/10.1111/j.2044-8317.2012.02049.x

Sheldrick, R. C., Frenette, E., Vera, J. D., Mackie, T. I., Martinez-Pedraza, F., Hoch, N., Eisenhower, A., Fettig, A., & Carter, A. S. (2019). What Drives Detection and Diagnosis of Autism Spectrum Disorder? Looking Under the Hood of a Multi-stage Screening Process in Early Intervention. *Journal of autism and developmental disorders*, *49*(6), 2304–2319. https://doi.org/10.1007/s10803-019-03913-5

Shim, S. R., Kim, S. J., & Lee, J. (2019). Diagnostic test accuracy: application and practice using R software. *Epidemiology and health*, *41*, e2019007. https://doi.org/10.4178/epih.e2019007

Simonoff, E., Pickles, A., Charman, T., Chandler, S., Loucas, T., & Baird, G. (2008). Psychiatric disorders in children with autism spectrum disorders: prevalence, comorbidity, and associated factors in a population-derived sample. *Journal of the American Academy of Child and Adolescent Psychiatry*, *47*(8), 921–929. https://doi.org/10.1097/CHI.0b013e318179964f

Statistics Solutions. (2025). Confirmatory Factor Analysis . Retrieved from https://www.statisticssolutions.com/academic-solutions/resources/directory-of-statistical-analyses/confirmatory-factor-analysis/

Tavakol, M., & Wetzel, A. (2020). Factor Analysis: a means for theory and instrument development in support of construct validity. *International journal of medical education*, *11*, 245–247. https://doi.org/10.5116/ijme.5f96.0f4a

Thabtah, F., & Peebles, D. (2019). Early Autism Screening: A Comprehensive Review. *International journal of environmental research and public health*, *16*(18), 3502. https://doi.org/10.3390/ijerph16183502

Toft, N., Jørgensen, E., & Højsgaard, S. (2005). Diagnosing diagnostic tests: evaluating the assumptions underlying the estimation of sensitivity and specificity in the absence of a gold standard. *Preventive veterinary medicine*, *68*(1), 19–33. https://doi.org/10.1016/j.prevetmed.2005.01.006

Uljarević, M., Spackman, E. K., Whitehouse, A. J. O., Frazier, T. W., Billingham, W., Condron, P., Hardan, A., & Leekam, S. R. (2023). Arriving at the empirically based conceptualization of restricted and repetitive behaviors: A systematic review and meta-analytic examination of factor analyses. *Clinical psychology review*, *103*, 102286. https://doi.org/10.1016/j.cpr.2023.102286

Van Steensel, F. J., Bögels, S. M., & de Bruin, E. I. (2013). Psychiatric comorbidity in children with autism spectrum disorders: A comparison with children with ADHD. *Journal of child and family studies*, *22*, 368-376.

Wieckowski, A. T., Williams, L. N., Rando, J., Lyall, K., & Robins, D. L. (2023). Sensitivity and Specificity of the Modified Checklist for Autism in Toddlers (Original and Revised): A Systematic Review and Meta-analysis. *JAMA pediatrics*, *177*(4), 373–383. https://doi.org/10.1001/jamapediatrics.2022.5975

Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior research methods*, *51*(1), 409–428. https://doi.org/10.3758/s13428-018-1055-2

Yaşlioğlu, M., & Toplu Yaşlioğlu, D. (2020). How and When to Use Which Fit Indices? A Practical and Critical Review of the Methodology. *Istanbul Management Journal*, 1–20. https://doi.org/10.26650/imj.2020.88.0001

Zeidan, J., Fombonne, E., Scorah, J., Ibrahim, A., Durkin, M. S., Saxena, S., Yusuf, A., Shih, A., & Elsabbagh, M. (2022). Global prevalence of autism: A systematic review update. *Autism research : official journal of the International Society for Autism Research*, *15*(5), 778–790. https://doi.org/10.1002/aur.2696