FACTOR ANALYSES AND CLINICAL DISCRIMINANT VALIDITY OF THE GILLIAM AUTISM RATING SCALE – 3RD EDITION (GARS-3) USING SPECIAL EDUCATION STAFF RATINGS IN SAMPLES WITH AUTISM SPECTRUM DISORDER AND OTHER DEVELOPMENTAL DISABILITIES

By

Nicole Bergamo Isbell

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

School Psychology – Doctor of Philosophy

2022

ABSTRACT

FACTOR ANALYSES AND CLINICAL DISCRIMINANT VALIDITY OF THE GILLIAM AUTISM RATING SCALE – 3RD EDITION (GARS-3) USING SPECIAL EDUCATION STAFF RATINGS IN SAMPLES WITH AUTISM SPECTRUM DISORDER AND OTHER DEVELOPMENTAL DISABILITIES

By

Nicole Bergamo Isbell

Based on Center for Disease Control and Prevention (CDC) surveillance reports, the prevalence of autism spectrum disorder (ASD) continues to increase (Maenner et al., 2021). As such, assessment tools that are efficient, cost-effective, and psychometrically sound are key to effective screening, accurate diagnosis, and clarification of intervention needs (Kuriakose & Shalev, 2016; Zwaigenbaum & Penner, 2018). The Gilliam Autism Rating Scale – Third Edition (GARS-3; Gilliam, 2013), a substantial revision from earlier editions, is a rating scale used to gather information from parents, caregivers, or teachers for screening or as part of a more comprehensive ASD assessment. Across editions, the GARS is considered a popular assessment tool among school psychologists (e.g., Aiello et al., 2017; Benson et al., 2019). However, despite the strong psychometric characteristics reported in the test manuals with standardization samples, prior editions were criticized for their screening performance in independent research samples, and factor analyses suggested problems with the test author's proposed subscales (e.g., Lecavalier, 2005; Pandolfi et al., 2010; South et al., 2002; Volker et al., 2016; Volker et al., 2022). To date, there has been little to no research focused on the psychometric properties of the current version of the GARS beyond what is reported in the test manual. Of critical importance, there have been no published independent factor analyses conducted in ASD or broader developmental disability samples and no independent estimates of screening effectiveness or clinical discriminant validity of the GARS-3. Therefore, the present project seeks to add to the limited research regarding the GARS-3 using program evaluation data from a large special education agency in Western New York state. The project consisted of three different studies that addressed aspects of GARS-3 internal structure validity and clinical discriminant validity. Study one involved an exploratory factor analysis (EFA) of the GARS-3 items with an ASD sample (*n* = 204) rated by special education teaching staff. Study two, confirmatory factor analyses (CFA) using a second ASD and non-ASD developmental disabilities (DDs) sample (*n* = 200), were used to examine the model fit of the published GARS-3 model and the factor model derived from the study one EFA, and assess which of the two models better fit the sample covariance matrix. Finally, aspects of the GARS-3's clinical discriminant validity were assessed using unique ASD cases from studies one and two (ASD sample *n* = 226) and an additional non-ASD developmental disabilities sample (non-ASD DDs sample *n* = 64) from the same special education agency. Clinical discriminant validity was examined via between-group comparisons, classification accuracy of a predetermined cut score, and exploration of other possible cut scores using receiver operator characteristic (ROC) curve analyses. The EFA resulted in a six-factor solution that was very similar in structure to the GARS-3 published six-factor model – differing only in the placement of one item. The CFAs indicated that the GARS-3 published model and the EFA-derived model both fit the data well and did not substantively differ. However, when cross-loadings were added, based on EFA results, CFA model fit significantly improved. ROC curve analyses indicated that, when using the suggested cut score of 70, sensitivity and specificity were lower than predicted. Lower cut scores yielded good sensitivity but poorer specificity, while higher cut scores showed the opposite pattern. Discussion and recommendations pertained to examining items and subscales based on cross-loadings and inter-factor correlations in addition to clinical implications of sensitivity and specificity findings.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Martin Volker, in addition to my dissertation committee members, Dr. Kristin Rispoli, Dr. Gloria Lee, and Dr. Connie Sung, for their help and mentorship. Thank you to my team of fellow researchers, Megan Stoll, Shelby Brennan, Dr. Jennifer Toomey, Dr. Amy Nasamran, and Dr. Nicole Mathes, for their expertise and support. Thank you to my dear friends and cohort-mates, Emma, Lake, and Rachel, for their daily encouragement throughout this process. I would also like to recognize my husband, Matt, my friends, and my family for their ongoing support. Thank you all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE: INTRODUCTION

Autism Spectrum Disorder (ASD)

Autism spectrum disorder (ASD) is a developmental disorder characterized by social communication and interaction difficulties as well as restricted interests and repetitive behaviors (American Psychiatric Association [APA], 2013). The current prevalence rate for ASD in the United States, based on 2018 surveillance data, is approximately one in 44 children (Maenner et al., 2021). Such prevalence estimates have gradually, and significantly, increased over the last twenty years (e.g., 1 in 54 according to data collected in 2016; 1 in 68 according to data collected in 2010; 1 in 150 from data collected in 2000), which may be due to a variety of factors (CDC, 2014; CDC, 2007; Maenner et al., 2020). However, the primary debate has centered around to what extent the changing numbers reflect true increases in actual cases of ASD, as opposed to more, already existing, cases being identified because of greater public awareness of ASD, broadened or differing definitions of ASD, and availability of improved assessment tools and intervention resources (Fombonne, 2003; Rutter, 2005). Regardless of the cause, as the prevalence rate increases, there is greater need for psychometrically sound (i.e., reliable, valid), cost-effective, and efficient assessment instruments for screening purposes and to aid in the diagnosis of ASD (Zwaigenbaum & Penner, 2018).

ASD Assessment

There are a variety of methods and tools used in the assessment of ASD. It is important for practitioners in multiple settings (e.g., school, clinical, medical) to be familiar with best practices, because it is imperative for early and accurate screening and diagnosis, as early intervention is clearly linked to better outcomes (Aiello et al., 2017; Kuriakose & Shalev, 2016). Typically, the first mode of assessment is screening (e.g., either level one or level two screeners for any atypical development or suspicion of ASD), and if needed, following-up with a comprehensive diagnostic assessment. Assessment is also linked to treatment and educational planning, monitoring progress toward intervention goals, and assessing treatment outcomes (Kuriakose & Shalev, 2016; Lord et al., 2005; Salvia et al., 2007).

Broadly speaking, best practice for comprehensive psychological, educational, and developmental assessments involve data collected via multiple methods across various sources (e.g., Pandolfi & Magyar, 2016). In the context of comprehensive diagnostic assessments for ASD, considering specific ASD diagnostic criteria, clinicians likely use semi-structured observation and interview techniques that are considered “gold standard” (e.g., the Autism Diagnostic Interview-Revised [ADI-R; Rutter et al., 2003b] and the Autism Diagnostic Observation Schedule – Second Edition [ADOS-2; Lord et al., 2012]). In addition, clinicians gather information using rating scales and other standardized assessments to measure domains of functioning critical to contextualizing an ASD diagnosis – this often includes the assessment of adaptive skills, language, and cognition (Kuriakose & Shalev, 2016). In schools, best practice assessment for special education eligibility is similar to a diagnostic assessment approach. This includes interviews with parents for developmental and medical history, assessment of current functioning by different measures including rating scales (e.g., social, communication, behavioral), direct and indirect observation of social interactions across contexts, and additional measures to assess associated domains (e.g., cognition, language, sensorimotor, adaptive; Aiello et al., 2017; Clark et al., 2014; Noland & Gabriels, 2004).

Gilliam Autism Rating Scale – 3rd Edition (GARS-3)

Rating scales are often an efficient way to collect data. They are a standardized way to obtain, compare, and contextualize information from multiple stakeholders (Norris & Lecavalier, 2010b; Scahill & Lord, 2004). Rating scales can be used as screening tools or to contribute information as part of a more comprehensive diagnostic assessment. Level one screeners can preliminarily assess for atypical development broadly or ASD-related symptomatology specifically. Level two screeners are used to assess for more specific symptoms after a level one screening tool has produced concerning results (Norris & Lecavalier, 2010b).

One ASD-specific rating scale is the Gilliam Autism Rating Scale – 3rd Edition (GARS-3; Gilliam, 2013). The GARS, across its various editions, is consistently reported among the top three most frequently used ASD assessment instruments in schools (GARS-2, Aiello et al., 2017; GARS-3, Benson et al., 2019). Results of a survey of school psychologists, conducted by Benson et al. (2019), indicated that the GARS-3 is widely used by school psychologists (i.e., almost 40% of their sample) and was the third highest used ASD-specific assessment tool, following closely behind the Childhood Autism Rating Scale – Second Edition (CARS2; Schopler et al., 2010) and the Autism Spectrum Rating Scales (ASRS; Goldstein & Naglieri, 2009). The GARS-3 is characterized as a level two screener and can be completed by parents, other caregivers, or teachers. The rating scale is comprised of 58 items that are distributed across six subscales. These subscales are Restricted/Repetitive Behaviors, Social Interaction, Social Communication, Emotional Responses, Cognitive Style, and Maladaptive Speech. All subscales yield norm-referenced scaled scores. An overall score, the Autism Index, is calculated by transforming the sum of subscale scores into a norm-referenced deviation quotient metric (i.e., normative *M* = 100, *SD* = 15). There are, however, two ways to calculate the Autism Index – depending on the communication status of the individual being rated. If the rated individual is communicative, the rater completes all six subscales, which are summed to form a composite score referred to as the Autism Index 6. If the individual is nonverbal/noncommunicative, the rater completes only the first four subscales, which are summed to form the Autism Index 4 composite. (See Appendix A [Tables A1 and A2] for a visual summary conceptualizing the GARS-3 subscales and composite scores.) Autism Index score ranges are connected to ordinal interpretive categories that indicate the probability of ASD – ranging from “unlikely” to “very likely” – in addition to the severity level. According to the manual, scores from the GARS-3 can be used to identify ASD, assess level of severity, indicate treatment progress, help to define educational or treatment goals, and as an ASD measure in research projects (Gilliam, 2013).

Internal Structure and Clinical Discriminant Validity

When choosing assessment methods to identify disorders such as ASD, it is important for researchers and practitioners to be familiar with the test’s psychometric properties (e.g., the normative sample, score characteristics, reliability, and validity). Internal structure validity and clinical discriminant validity are both aspects of broader construct validity and are of particular importance in this project. Specifically, an assessment’s factor structure (i.e., internal structure) should provide support for the instrument’s scores and their construct interpretations (Brown, 2015; Floyd & Widaman, 1995). Additionally, the assessment should be able to accurately discriminate between cases known to be at different levels of the intended construct measured by the instrument. In the case of the GARS-3, the instrument should accurately differentiate between individuals with and without ASD (Kuriakose & Shalev, 2016).

As a statistical method, factor analysis can be used to examine underlying constructs (Floyd & Widaman, 1995; Gorsuch, 1983). Specifically, in the rating scale context, exploratory factor analysis (EFA) seeks to explain inter-item correlations through a smaller number of latent variables (i.e., constructs). Confirmatory factor analysis (CFA) also seeks to explain covariation among observed variables (e.g., items from a rating scale) in terms of the influence of a smaller number of latent variables. However, in the case of CFA, the model – consisting of observed variables, latent variables, and the relationships among them – is specified prior to the analysis (based on theory or prior empirical findings) and then assessed for degree of fit with the variance-covariance matrix of the observed variables (Brown, 2015; Floyd & Widaman, 1995). Confirmatory factor analysis is especially useful in validating an assessment instrument. Using CFA, researchers can assess how well a model, based on the scoring structure for the test, fits the observed relationships among the test items, and how generalizable that model fit is across different samples (Floyd & Widaman, 1995). Such samples may reflect differences in sampling error, age, ethnicity, diagnostic status, rater types, or other potentially important variables that could impact the relationships among test items.

Clinical discriminant validity is also important in determining the utility of a scale. An instrument used to screen for a particular disorder, such as the GARS-3 with ASD, should have the ability to discriminate individuals with a higher likelihood of having ASD from those who have a low likelihood. Sensitivity and specificity are indices of how effective an instrument is at this kind of differentiation. Sensitivity reflects how well a measure like the GARS-3 accurately identifies individuals with a specific diagnosis (e.g., ASD), while specificity looks at how well the measure identifies individuals that do not have the diagnosis (Bandalos, 2018; Lalkhen & McCluskey, 2008). ASD has specific, core diagnostic features (APA, 2013), though an individual may have a different diagnosis that presents with similar or partially overlapping symptomatology. This may make it difficult to discriminate between different types of samples (e.g., between ASD and other developmental disabilities [DDs]; Volker et al., 2016). Therefore, it is important for an assessment measure to have the psychometric evidence to support its use in identifying a disorder *and* in discriminating that disorder from other disorders.

Purpose of the Study

The purpose of this study is to address the gap in the psychometric literature for the GARS-3 using ratings from special education staff for samples of individuals with ASD. The GARS-3, and its prior versions, is among the most widely used ASD-related assessment tools in the school setting (e.g., Aiello et al., 2017; Benson et al., 2019). However, its factor structure has not been independently evaluated. Independent research with prior editions of the GARS found results that were discrepant with the factor structure reported in the manual – both in terms of the number of factors retained (e.g., Pandolfi et al., 2010) and, in all studies, in terms of item assignments to factor-based subscales (e.g., Lecavalier, 2005; Volker et al., 2016; Volker et al., 2022). Further, there is no published independent research regarding the GARS-3's ability to accurately classify individuals with ASD. Previous research with the GARS and GARS-2 suggested that with the recommended cut score, there were high rates of false negatives in ASD research samples (e.g., South et al. [2002] reported 52%; Volker et al. [2016] reported 34.7%). Given these findings of poor factor structure generalizability and poor clinical discriminant validity for prior versions of the instrument in independent samples, it is imperative that the psychometric properties of the newest edition – the GARS-3 – be examined in additional samples and with different rater types who are likely to use it in practice.

This dissertation contributed to the literature by providing evidence for the internal structure and clinical discriminant validity for the GARS-3 in a series of three studies. Study one involved an EFA of the GARS-3 items rated by special education teaching staff within a sample of students with ASD. Study two involved two CFAs examining the fit of Gilliam's (2013) proposed six subscale model and the model based on the EFA in study one. Additionally, the two models were compared to determine a better fit with the data. These CFAs were conducted using a different sample (separate from study one) consisting of individuals with ASD and non-ASD DD diagnoses. In study three, clinical discriminant validity was examined between two diagnostic groups: ASD and non-ASD DDs. Mean differences in GARS-3 composite scores, classification accuracy (e.g., sensitivity and specificity) of the cut score recommended in the GARS-3 manual, and the utility of possible alternative cut scores were examined.

In the following chapters, this paper will explore ASD, assessment strategies, and methods of validating assessment instruments. Specifically, chapter two (Literature Review) will illustrate ASD diagnostic criteria from a historical lens, in addition to examining commonly comorbid features that may add complexity to ASD assessment. This review will provide information about ASD assessment techniques and contextualize the use of rating scales. Further, this chapter will explore the purposes of factor analytic procedures and their use in validating rating scales, in addition to methods to support the clinical discriminant validity of an assessment tool. Specific to the GARS-3, this review will provide information about the three different versions of the scale, and evidence from the literature that contests the GARS and GARS-2 factor structure suggested from the manual. The third chapter (Method) will provide rationales for each study's research question(s) in addition to specific information about the samples and analyses for the three studies (i.e., the exploratory and confirmatory factor analyses and clinical discriminant validity). The fourth chapter (Results) will present the findings of each of the three studies through the eight research questions and accompanying hypotheses. Finally, the fifth chapter (Discussion) will summarize and provide context for the results in addition to exploring strengths, limitations, and implications of the study and its findings.

CHAPTER TWO: LITERATURE REVIEW

ASD Diagnostic Criteria

The term “autism” has developed from being a specific symptom to being conceptualized as a spectrum disorder. In the early 1900s, Eugen Bleuler used “autism” to describe a symptom of schizophrenia and indicated that “autistic thinking” was a way to avoid reality (Evans, 2013). Around the 1940s, Leo Kanner provided his own conceptualization of autism. He characterized “early infantile autism” as a separate disorder from schizophrenia. This included diagnostic features such as deficits in social interaction, language, and communication as well as rigidity (Kanner, 1942; Volker et al., 2012). Future research focused on different aspects of Kanner’s case descriptions. For example, in 1967, Bruno Bettelheim emphasized the psychoanalytic perspective that autism occurred as a defense mechanism to unavailable parents/parenting techniques. In contrast, in 1964, Bernard Rimland examined autism through a biological lens and considered it a genetic and neurodevelopmental disorder (Volker et al., 2012).

Although autism was not officially listed until the third edition of the Diagnostic Statistical Manual (DSM-III; APA, 1980), cases of autism were likely captured in other diagnoses such as childhood schizophrenia in the DSM (APA, 1952) and DSM-II (APA, 1968). As a pervasive developmental disorder (PDD), infantile autism made its first appearance in the DSM-III (APA, 1980) and was later grouped with additional PDDs; first, with pervasive developmental disorder – not otherwise specified (PDD-NOS) in the DSM-III-R (APA, 1987), and then with three additional PDDs in the DSM-IV (APA, 1994) and DSM-IV-TR (APA, 2000; Volker et al., 2012). Currently, in the DSM-5 (APA, 2013), autism is conceptualized as one disorder on a spectrum – to be explored in more detail in the following sections. Across different iterations of the DSM, both revisions and new editions, the way autism is conceptualized has evolved. However, core features of social and communication difficulties in addition to restricted behaviors, identified by early researchers such as Kanner, have largely remained consistent to the broad diagnosis.

DSM-IV-TR Diagnostic Criteria

In the DSM-IV (APA, 1994) and the DSM-IV-TR (APA, 2000), five diagnoses were listed as pervasive developmental disorders: autistic disorder, Asperger's disorder, Rett's disorder, childhood disintegrative disorder (CDD), and pervasive developmental disorder—not otherwise specified (PDD-NOS). Asperger's and PDD-NOS were considered to have milder symptom severity while CDD involved extreme developmental regression. Rett's disorder, predominantly found in females, was characterized with difficulties in movement and communication. In order to meet diagnostic criteria for autistic disorder, children must exhibit at least six different behavioral criteria across three categories. Within these categories, at least two behaviors must be noted from category A.1 (i.e., impairments in social interaction), one behavior from category A.2 (i.e., impairments in communication), and one behavior from category A.3 (i.e., restricted repetitive and stereotyped behaviors, interests, and activities). Rule out diagnoses included another PDD that better captured or explained the exhibited behaviors. Additionally, the DSM-IV-TR specified that autistic disorder could not be comorbidly diagnosed with attention-deficit/hyperactivity disorder (ADHD) or stereotyped movement disorder (APA, 2000; Harker & Stone, 2014).

DSM-5 Diagnostic Criteria

In the fifth edition of the DSM (APA, 2013), there were marked changes in how autism was conceptualized. Autism spectrum disorder (ASD) emerged as a diagnosis that subsumed previous diagnoses of autistic disorder, Asperger's, and PDD-NOS (APA, 2013). Rett's disorder and CDD, previously viewed as PDDs, were removed from this category (Harker & Stone, 2014). The DSM-IV-TR characterized autism with three categories: impairments in social interaction, communication, and restricted and repetitive behaviors. In the DSM-5, ASD is conceptualized in *two* categories: category A, which pertains to deficits in social communication and social interaction and category B regarding restricted and/or repetitive patterns of behavior, interests, or activities. To warrant a diagnosis, the three behavioral criteria in category A must be met. This includes difficulties with social-emotional reciprocity (e.g., reciprocal conversations, difficulties initiating social interactions, limited sharing of interests, affect, and/or emotions), deficits in nonverbal communication (e.g., lack of eye contact, limited use of gestures and facial expressions), and deficits in developing, maintaining, and understanding relationships (e.g., difficulties being flexible in different social interactions, difficulties with imaginary play). Additionally, at least two behavioral criteria in category B must be met. Examples include stereotyped or repetitive motor movements, use of objects, or speech (e.g., lining up toys, echolalia), inflexibility (e.g., ritualized behaviors, marked distress when something changes), restricted interests (e.g., strong preoccupation with an unusual object/interest or unusually intense preoccupation with a specific object/interest), and sensory difficulties (e.g., hyper- or hypoactivity to sensory input or interests). Regarding rule out diagnoses indicated that difficulties should not be better explained by an intellectual disability or a global developmental delay. Further, comorbidities, such as ADHD, are allowed and recognized, but specifiers for diagnosing ASD should be indicated (i.e., with or without intellectual impairment, language impairment, a known medical or genetic condition or environmental factor, with another neurodevelopmental, mental or behavioral disorder, or with catatonia; adapted from DSM-5, APA, 2013, p. 50-51; Harker & Stone, 2014).

Common Comorbidities

There are many disorders that frequently comorbidly occur with ASD. Selected literature from the DSM-5 estimates that 70% of individuals with ASD have one comorbid disorder and 40% have more than one comorbidity. Common comorbid disorders include intellectual disability, speech or language issues, seizures or epilepsy, ADHD, as well as other psychiatric disorders. Researchers have also observed associated features of difficulties with sensory input and specific characteristics and patterns associated with gender (APA, 2013).

Intellectual Disability (ID)

Intellectual disability (ID) and ASD are commonly comorbid, with ID more frequently occurring with ASD compared to other diagnoses (APA, 2013). In previous versions of the DSM, ID was conceptualized as mental retardation (MR; Volker et al., 2012). The term intellectual disability/intellectual developmental disorder was introduced to the DSM-5 in 2013 (APA, 2013). Historically, prevalence estimates have indicated that the majority of ASD cases also have comorbid ID. For example, Shea and Mesibov (2005) reported 70-80% of individuals with autistic disorder were reported with comorbid MR, and Matson and Shoemaker (2009) reported 50-70% of individuals with ASD were also diagnosed with ID (Peters-Scheffer et al., 2016; Volker et al., 2012). Current prevalence estimates from the CDC indicate that only 33% of the sample had ID – although only some cases had information about cognition/intelligence – and more girls were identified with comorbid ASD and ID compared to boys (Maenner et al., 2020)

ID involves deficits in intellectual abilities (e.g., IQ below 70) and adaptive functioning, overlapping in select symptomatology with ASD (APA, 2013; Peters-Scheffer et al., 2016). Specifically, individuals with ASD or ID may have difficulties with social interactions and speech (e.g., delayed, nonverbal) and may also have stereotyped motor movements (Peters-Scheffer et al., 2016). Individuals with comorbid ASD and ID might have greater difficulties with social interactions and adaptive skills and have higher rates of additional comorbid disorders compared to individuals with a sole diagnosis of ID (e.g., psychosis, ADHD, internalizing disorders; Matson & Shoemaker, 2009). Notably, research has shown that the higher severity level of ID correlates with greater ASD symptom severity (Matson & Shoemaker, 2009; Peters-Scheffer et al., 2016).

This comorbidity may be difficult to diagnose, especially in young children or when individuals with ID have little to no intact language or means of communication. In assessment, it is therefore important for a clinician to gather data regarding both verbal and nonverbal abilities (APA, 2013). Although both disorders are associated with social deficits, ID is comorbid with ASD when difficulties with social communication fall below what is expected for the individual's developmental level (APA, 2013; Matson & Shoemaker, 2009). There may also be overlapping criteria for the two diagnoses such as the presence of repetitive behaviors (APA, 2013). Peters-Scheffer et al. (2016) indicate several screening instruments developed for this comorbidity including the Social Communication Questionnaire (SCQ; Rutter et al., 2003a) and the Autism Behavior Checklist (ABC; Krug et al., 1980). Although ASD and ID can have overlapping features, it is important to separate differences for accurate diagnosis and if they do occur together, to attend to unique challenges of this comorbid pair.

Speech and Language Difficulties

Language – one part of social communication and interaction – is a core difficulty for individuals with ASD. However, when addressing a language impairment, diagnosticians define this specifically as a structural language abnormality at least 1.5 standard deviations below the mean of a standardized assessment (Boucher, 2012). Individuals with ASD may have no language impairments, or, mild to severe language impairments (Boucher, 2012).

Research indicates trends associated with language impairments of individuals on the spectrum such as a common occurrence of delayed language. This might include difficulties with comprehension, "odd" phrases, and/or atypical grammar and articulation. It is reported that school-age children with ASD tend to have the most difficulty with comprehension, semantics, and specific morphology (Boucher, 2012). Additional research has examined ASD symptom severity related to language impairment. A study by Loucas and colleagues (2008) did not find a correlation between these variables but did find that those with ASD without language impairment had greater social adaptation – although skills were still below average. Additionally, results indicated that individuals with ASD and a language impairment had greater difficulties with functional communication and receptive language compared to non-ASD individuals with a language impairment (Loucas et al., 2008).

In assessment for ASD, a clinician should provide clear descriptions of the individual's speech and assess both receptive and expressive language. When specifying a language impairment, a clinician might indicate that the individual has "no intelligible speech (nonverbal), single words only, or phrase speech" (APA, 2013, p. 53). In all, a language impairment adds complexity to the ASD profile, beyond deficits in social communication, which should be specified in the diagnosis.

Seizure Activity/Epilepsy

Epilepsy is a disorder characterized by recurrent seizures and is more prevalent in the ASD population compared to the general population (e.g., 30% versus 2-3%; Boothe & Zuna, 2019; Tuchman et al., 2010). The literature suggests two "peaks of onset" of epilepsy for individuals with ASD, either in early childhood or adolescence (Boothe & Zuna, 2019; Tuchman et al., 2010). These comorbid disorders are found in individuals regardless of the level of ASD symptom severity; however, research has higher prevalence associated with ID (APA, 2013; Boothe & Zuna, 2019; Tuchman et al., 2010). Research also suggests this comorbid pair is associated with difficulties with adaptive behavior, communication and/or pragmatics, motor functioning, and behavior (e.g., difficulties with attention, hyperactivity, mood, aggression; Boothe & Zuna, 2019; Tuchman et al., 2010). A diagnosis of epilepsy or presence of seizures in addition to ASD may introduce complexities in assessment (e.g., associated behavioral difficulties) and treatment (Boothe & Zuna, 2019).

ADHD

Attention deficit/hyperactivity disorder (ADHD) is commonly comorbid with ASD. Prevalence estimates can widely range from 14% to 78%, with several studies estimating above 50% (Jang et al., 2013; Stevens et al., 2016). Prior to the DSM-5, psychologists could not comorbidly diagnose ADHD in individuals with ASD (Antshel & Russo, 2019; Jang et al., 2013). While both disorders may characteristics of social and executive functioning difficulties, ADHD should be made as a differential and comorbid diagnosis when these difficulties are above and beyond what would be expected of an individual of the same mental age (APA, 2013; Antshel & Russo, 2019). ADHD is specifically characterized by inattentive behavior (e.g., inattention to details or instructions, difficulty with organization and remaining focused) and/or hyperactive/impulsive behavior (e.g., frequent fidgeting, difficulty waiting for their turn). These difficulties often lead to clinical impairments in the individual’s academic and social life (APA, 2013, pp. 59-63).

Research indicates associated trends in these comorbid disorders and there is evidence to suggest influences on assessment. A study by Stevens et al. (2016) found that individuals with comorbid ASD and ADHD – compared to groups with ASD only, ASD + ID, and ASD + ADHD + ID – were significantly older when parents first noticed potential developmental problems, when parents sought guidance for concerns from a doctor, and when they received an ASD diagnosis. Since ASD and ADHD have some overlapping core features, clinicians should look critically at information gathered. Parents and teachers who provide the needed information to contribute to a diagnosis, might associate behaviors with specific disorders and chosen assessments might lack discriminant validity (e.g., measuring features of different disorders using similar constructs; Antshel & Russo, 2019; Stevens et al., 2016). For example, the Social Communication Questionnaire (SCQ) has been used to differentiate between ADHD and ASD, however, there is evidence to suggest that gold-standard ASD measures such as the ADOS and ADI-R contain items with low discriminant abilities (Antshel & Russo, 2019).

Other Psychiatric Comorbidities

In the DSM-5, one or more emotional and/or behavioral disorders can be comorbidly diagnosed with ASD, although there is debate about some co-occurrence of disorders (Pandolfi & Magyar, 2016). In a literature review identifying papers which examined ASD and comorbid disorders since 2000, researchers found that the majority of papers examined ADHD, followed by anxiety and depression (Matson & Cervantes, 2014). Anxiety and mood disorders are the most common co-occurring emotional disorders with ASD and prevalence estimates from the literature range from 11-84% for anxiety disorders and 1-38% for mood disorders. The most common diagnoses were specific phobias, obsessive compulsive disorder (OCD) and social anxiety disorder. Regarding mood disorders, among the most common were depressive disorders and bipolar disorder (Pandolfi & Magyar, 2016). Other comorbid disorders and difficulties studied in the literature include oppositional defiant disorder, conduct disorder, tics, schizophrenia, mania, psychosis, somatic difficulties, post-traumatic stress, and disorders involving feeding, elimination, and/or sleep (APA, 2013; Matson & Cervantes, 2014). The DSM-5 lists many of these disorders co-occurring with ASD, in addition to developmental coordination disorder (APA, 2013).

As mentioned in other sections, comorbidities with ASD contribute challenges in assessment and conceptualization. Clinicians, especially those who may be less experienced, may have greater difficulty separating ASD symptomatology from other psychiatric features (Matson & Cervantes, 2014). Moreover, common difficulties with communication or with cognition/processing may make it more difficult for the individual with ASD to describe internalizing difficulties. Assessment should identify ASD and other psychiatric features through best practice methods (e.g., multi-method and multi-source data collection). However, there is a lack of psychometrically strong assessment tools to accurately identify emotional disorders with individuals with ASD (Pandolfi & Magyar, 2016).

Sex Differences

Research has noted sex differences within ASD both in prevalence of diagnoses and symptom presentation. The latest prevalence estimate indicated a male to female ratio of 4:1. This pattern of higher rates of ASD in males has generally remained the same since 2002 (Maenner et al., 2020). Research has also shown a trend of higher prevalence of ASD in females who have comorbid ID (APA, 2013).

Research on sex differences within the ASD population has been mixed regarding core feature presentation. Generally, research does not indicate sex differences in ASD, but some studies have shown that females may have fewer restricted and repetitive behaviors, but increased difficulties with social affect, motor functions, adaptive behavior, and emotion (Matheis et al., 2019). Research by Matheis and colleagues (2019) found no sex differences in a sample of 1,317 children ages 17-37 months. However, they did find that there were significant differences when grouped by the presence or absence of a cognitive delay. Regarding developmental functioning, researchers found sex differences such that females presented with greater motor difficulties and males had greater difficulties with communication (Matheis et al., 2019). In a study by Nasca et al. (2019), researchers found, in their sample of individuals with ASD (without ID) matched for IQ, no significant sex differences in parent-reported externalizing and internalizing difficulties.

In all, there seems to be mixed findings regarding sex differences. These differences might be due to biological factors but could also be due to assessment and diagnosis. Females without ID or language difficulties may be less likely to be diagnosed or diagnosed later – possibly due to different and more subtle presentations in social and communication challenges, and/or because they are better able to mask their social difficulties (APA, 2013, Matheis et al., 2019). Moreover, much of the historical research (e.g., Kanner) has been focused on males. Assessment measures may also lack the sensitivity in identifying ASD in females as research indicates gender differences in ADI-R and ADOS scores (Matheis et al., 2019). Overall, sex differences are important to consider regarding diagnosis, characteristics, and development. The current research reflects higher prevalence in males, but also mixed findings related to symptomatology differences (Lai et al., 2015; Matheis et al., 2019).

Sensory Processing Difficulties

Many individuals on the spectrum have hypersensitivity, hyposensitivity, and/or atypical responses to sensory stimuli (Ashburner et al., 2008). Specifically, individuals may have differences, compared to typically developing (TD) peers, regarding auditory and visual processing, and to tactile input (Tomchek & Dunn, 2007). The research suggests that individuals with ASD may have a relative strength processing simple sensory input, but greater difficulty with complex stimuli. Researchers hypothesize that individuals may seek sensory input when stressed or under stimulated, and may favor predictable versus unpredictable input (e.g., favor mouthing familiar objects, but disliking new foods; Ashburner et al., 2008).

Sensory input can be studied using observation, retrospective recordings, and parent and self-report. One tool, the Short Sensory Profile, has been used in the literature to capture sensory differences, and measures different aspects of sensitivity: tactile sensitivity, taste/smell sensitivity, movement sensitivity (e.g., dislikes when feet leave the ground), under responsive/seeks sensations, auditory filtering (e.g., doesn’t respond or appears to not have heard something), low energy/weak (e.g., trouble physically supporting themselves), and visual/auditory sensitivity. Research using this measure has shown that the majority of an ASD sample of young children (i.e., 95%) had an impairment in sensory processing (Tomchek & Dunn, 2007). Additional research indicated poorer educational outcomes (e.g., achievement) associated with sensory difficulties (e.g., difficulty understanding verbal instructions in a louder setting, presence of sensory-seeking behaviors). Further, this study did not find a relationship between sensory processing and IQ (Ashburner et al., 2008).

Comorbidities and Associated Features in Assessment

These common comorbidities add obvious complexity to assessment. While details of assessment are explored more in the following section, it is important to consider comorbid presentations within different measures and assessment batteries. Specifically, it is necessary to critically examine assessed constructs and separate core and associated features of ASD. No assessment measure will likely be able to measure all such features. However, when using tools in screening or assessment, a clinician should be familiar with the tools' measured constructs – potentially identifying any missing constructs – and the reliability and validity of the measure.

Assessment of ASD Diagnostic Criteria

Purposes of ASD Assessment

Assessment with individuals with suspected or diagnosed ASD have a range of purposes that include screening, diagnosis, progress monitoring, outcome assessment, and identifying associated or comorbid features. Typically, a doctor or pediatrician is a family's first interaction with ASD screening where best practice indicates the use of developmental screeners. Screeners may indicate the possible need for follow-up assessment. Early identification in assessment is imperative, especially for children with ASD, as early intervention is linked to positive future outcomes. Research also indicates that ASD can be reliably diagnosed by two years of age (Kuriakose & Shalev, 2016).

However, there are also many barriers to assessment as differences have been reported due to IQ, age, gender, ethnicity, parent education, socioeconomic status, and geographic location. Although reliable diagnoses can be made at two years of age, the most recent CDC surveillance report indicated the median age of diagnosis reported by sites ranged from 29 to 46 months, and those reported in diagnostic reports ranged from 39 to 57 months. This data also indicated that children with lower IQs (e.g., below 70) were assessed early and of those individuals, Black children had a higher median age of diagnosis (Maenner et al., 2020). Several individual, family, and community factors may also influence the age of diagnosis. For example, factors associated with an earlier diagnosis include children who are male or have lower IQs and developmental regression, children with higher parent education and income levels, and children who live closer to medical centers. In contrast, children who have lower levels of socioeconomic status, who are ethnic minorities, and who live in rural settings have lower rates of or later diagnoses (Kuriakose & Shalev, 2016). Overall, it is important for clinicians to screen patients while noting patterns of individuals who may have missed or late diagnoses.

Screening

Screening is typically the first means of identifying concerns or developmental delays associated with ASD. Information is usually provided by a parent or caregiver in the form of a rating scale (Kuriakose & Shalev, 2016). The American Association of Pediatrics (AAP) recommends using ASD-specific screening instruments at 18 and 24 months (Hymen et al., 2020). Important psychometrics for screeners, similar to other rating scales, include their sensitivity (i.e., accurately identifying children who are at risk for ASD) and specificity (i.e., correctly identifying children who are *not* at risk for ASD). Ideally, rates of sensitivity and specificity exceed .80 (Kuriakose & Shalev, 2016). Screeners can be categorized as level one or level two screeners, depending on purpose of the assessment.

Level One Screening. In general, level one screeners are meant to identify children with any atypical development, or who may be at risk for developmental disorders and language or motor delays (Norris & Lecavalier, 2010b). Level one screeners for ASD seek to identify ASD-specific concerns or delays. The assessments are short to fill out, score, and interpret (Kuriakose

& Shalev, 2016). Screeners are designed to be highly sensitive in order to identify the greatest number of children who are at-risk. However, low specificity may lead to high rates of false positives in over-identifying individuals (Kuriakose & Shalev, 2016). One example of a level one screener is the M-CHAT-Revised with Follow-Up (M-CHAT-R/F; Robins et al., 2009). This scale is an extension of the first developed and validated level one screener, the Checklist for Autism in Toddlers (CHAT; Baron-Cohen et al., 1992; Kuriakose & Shalev, 2016).

M-CHAT-R/F. The M-CHAT-R/F is classified as a level one screener for ASD, intended for toddlers between 16 and 30 months. It was originally designed to assess risk for ASD or other developmental disabilities. In this revised version, parents rate 20 items – formatted as yes/no questions – regarding specific behaviors. One numerical score is calculated from the 20 items which corresponds to categories for the level of risk for ASD (i.e., low-, medium-, or high-risk). For those who fall in the "medium" or "high" risk categories, there are follow-up questions that gather more information pertinent to next steps (e.g., potential referral for diagnostic evaluation; Robins et al., 2009; Robins et al., 2014). These scores were standardized on a sample of 16,115 "low risk" toddlers in the US. Researchers reported adequate internal consistency reliability (Cronbach's α = 0.79). They also reported sensitivity of .85 and specificity of .99 (Robins et al., 2014).

Overall, the M-CHAT-R/F has advantages in ease and efficiency of administration in addition to being an early identification tool for this young population. However, research has also shown that there are high levels of false positives, the follow-up questions may be time-consuming, and there is limited psychometric evidence for the measure (Robins et al., 2009; Robins et al., 2014).

Level Two Screening. Level two screeners are used to assess children who are already identified as “at-risk” (e.g., from a level one screener; Norris & Lecavalier, 2010b). They are used to follow-up with families after collected information indicated that further assessment was necessary. While level one screeners are highly sensitive, level two screeners prioritize specificity – ruling out individuals who are *not* at-risk for an ASD diagnosis or to aid in determining differential diagnoses (Kuriakose & Shalev, 2016). Commonly used level two screeners include the Social Communication Questionnaire (SCQ; Rutter et al., 2003a), Autism Spectrum Rating Scales (ASRS; Goldstein & Naglieri, 2009), Social Responsiveness Scale – Second Edition (SRS-2; Constantino & Gruber, 2012), Childhood Autism Rating Scale – Second Edition (CARS2; Schopler et al., 2010), and Gilliam Autism Rating Scale – Third Edition (GARS-3; Gilliam, 2013; Benson et al., 2019; Kuriakose & Shalev, 2016). It is important to note that while these rating scales are often used as level two screeners, certain scales – like the CARS2 and the GARS-3 – are also used to contribute to diagnostic assessment, although not in place of “gold standard” diagnostic assessment tools (Kuriakose & Shalev, 2016).

SCQ. The SCQ is intended to be used as a screening tool to assess for ASD-related symptomatology (Rutter et al., 2003a). The SCQ is a parent-reported measure meant for children with a chronological age above four years and a mental age above two years. It includes 40 yes/no questions and has two different forms (Rutter et al., 2003b). The Lifetime form assesses lifetime behavior and obtains information similar to what is needed for a diagnosis, while the Current form assesses current behavior and contributes information used for treatment and education planning (Rutter et al., 2003a; Rutter et al., 2003b). There are four subscales: Social Interaction, Communication, Abnormal Language, and Stereotyped Behaviors). The Lifetime form has a total score that is compared to a cutoff score to indicate potential need for future assessment (Rutter et al., 2003b).

The standardization sample for this measure primarily consisted of individuals with autism (i.e., 160 of 200 individuals). Authors reported internal consistency reliability of the Total Score (i.e., alpha coefficient of .90), concurrent validity (e.g., high correlation between the Total Score and the ADI-R), and construct validity (e.g., indicating a four-factor solution). Authors reported a sensitivity of .85 and specificity of .75 when differentiating between individuals with ASD from other diagnoses (Berument et al.,1999). Authors also suggest using different cut scores when differentiating ASD from individuals without ASD or from individuals with other developmental disorders (Barnard-Brak et al., 2016; Corsello et al., 2007; Rutter et al., 2003a; Wiggins et a al., 2007).

Some of the benefits of the SCQ are its ease in scoring and administration in addition to demonstrating good reliability and validity. Some research suggests limitations include difficulty discriminating ASD from other developmental disorders, adjustment of cut scores depending on the population, and in identifying difficulties in children under a chronological age of four years or mental age of two years (Chandler et al., 2007; Corsello et al., 2007; Rutter et al., 2003a; Wiggins et al., 2007; Witwer & Lecavalier, 2007). Overall, it should be noted that the SCQ is designed as a screener to indicate the need for further evaluation.

ASRS. The ASRS can be used to measure ASD-related symptoms with individuals aged 2-18 years for screening (e.g., with the short form) and as part of a more comprehensive ASD evaluation (e.g., with the full-length form). Parents, teachers, and/or other caregivers rate the frequency of a child or adolescent’s behavior for the past four weeks. The scales consist of 70 or 71 items which are rated on a rating scale ranging from zero (never) to four (very frequently).

Items pertain to peer and adult socialization, social/emotional reciprocity, atypical language, stereotypy and behavioral rigidity, sensory difficulties, attention, and self-regulation. The short form yields one *T*-score while the full-length form reports four scores: Total, ASRS, DSM-IV-TR, and Treatment (Goldstein & Naglieri, 2013).

The measure was standardized with 2,560 individuals who predominantly did not have a clinical diagnosis, but also included individuals with autism and other diagnoses (e.g., ADHD, developmental delay; Goldstein & Naglieri, 2013). Test authors reported information on reliability: high internal consistency estimates (e.g., Cronbach's α = .97 for Total Score), test-retest reliability (i.e., for all scales, ranging .70 - .93), and interrater reliability (e.g., for all scales, ranging .73 - .92 for parent reports and .59 - .73 for teacher reports; Simek & Wahlberg, 2011). Additionally, test authors reported the validity of the ASRS. The ASRS is highly correlated with the GARS-2 and Gilliam Asperger's Disorder Scale (GADS; Gilliam, 2001) – other parent and teacher rated measures. However, it had lower correlations with the CARS, a clinician-rated measure (Kluck, 2014). Test authors reported high sensitivity and specificity (Goldstein & Naglieri, 2013).

Benefits of the ASRS include the availability of two forms of different lengths (found to be highly correlated with each other: r = .84 to .92), its efficiency to complete, and its data showing good reliability and validity. Several shortcomings of the measure include that the normative sample is not specifically ASD and the presence of potential difficulties discriminating between ASD and ADHD with specific scales (e.g., Self-Regulation, Attention; Goldstein & Naglieri, 2009; Kluck, 2014; Shaw, 2014; Simek & Wahlberg, 2011). Additionally, Shaw (2014) proposed that the ASRS may not contribute significant data beyond information directly collected by a clinician.

SRS-2. The SRS-2 is used to identify ASD-associated social-communication deficits through ratings typically from parent and teacher informants. The SRS-2 consists of 65 items with different forms depending on age (Preschool form for 2.5-4.5-year-olds, School-Age form for 4-18-year-olds, and an Adult form). Raters indicate the frequency of a behavior on a rating scale ranging from one (not true) to four (almost always true). There is one overall composite score, five treatment subscales, and two DSM-5 subscales, all reported as *T*-scores (Constantino & Gruber, 2012). This brief review will focus on the School-Age form as this is the original form with the most research and includes the most similar age range to other ASD rating scales.

In the normative sample, clinical diagnoses were not reported (Hoff & Doepke, 2014). Test authors indicated high internal consistency reliability (α = .94 - .96 for all ages in the total sample) and good interrater reliability (*r* = .61 for total sample; Constantino & Gruber, 2012). Authors reported moderate to high correlations with other ASD rating scales (e.g., SCQ and CARS) and low to moderate correlations with ASD diagnostic tools (e.g., ADI-R, ADOS; Bruni, 2014). They also reported good discriminate validity with high levels of sensitivity and specificity (.92) in addition to good construct validity with a CFA indicating a two-factor structure (Bruni, 2014; Constantino & Gruber, 2012).

Overall, the SRS-2 has adequate reports of reliability and validity, particularly with the School-Age form. There are also versions with extended age ranges (e.g., preschool and adult ages). The SRS-2 protocol does not label the instrument as a measure for ASD – possibly decreasing skewed perspectives/ratings based on this label. However, some limitations of this measure include unreported diagnoses in the standardization sample, no reported test-retest reliability (Bruni, 2014; Constantino & Gruber, 2012), and additional factor analytic research that yielded a four-factor solution – not aligned with the test authors' findings (Nelson et al., 2016).

CARS2. The CARS2 is a unique rating scale such that its main forms are rated by the clinician, not the parent (Schopler et al., 2010). It is intended to inform diagnostic evaluation when there is a high likelihood of an ASD diagnosis. There are two clinician-rated forms that consist of 15 items which use information from observation and/or interview. Items are rated on a scale ranging from one (behavior is within normal limits) to four (behavior is markedly abnormal or severe compared to same-age peers; Vaughan, 2011). Items measure ASD features such as emotions and emotional control, communication, restrictive and repetitive behaviors and interests, and social communication and interaction. The CARS2-High Functioning Version (CARS-HF) is intended for individuals above the age of six who have an IQ above 80 and intact language. The CARS2-Standard Version (CARS2-ST) is intended for individuals under the age of six or individuals above six who have IQ and language lower than the standards for the CARS-HF. (Schopler et al., 2010) In both the CARS2-ST and -HF, items produce a Total score which is compared to a cut score (Schopler et al., 2010; Vaughan, 2011).

The standardization sample for the CARS2-ST had IQs below 85 and an ASD diagnosis, while the sample for the CARS2-HF included those with an IQ of at least 80 and with diagnoses of ASD, other disorders (e.g., ADHD, learning, etc.), and students with or without special education services. Authors reported internal consistency reliabilities of .93 (CARS2-ST) and .96 (CARS2-HF) for total scores and interrater reliability of .95 for the CARS2-HF total score. Authors reported no other reliability information for the updated versions but indicated that the original CARS had interrater reliability ranging from .73-.83 and test-retest reliability of .88 (Schopler et al., 2010; Vaughan, 2011). Regarding validity, authors found moderate to high internal consistency, moderate correlations with SRS parent ratings, high correlations with the ADOS and ABC, and high sensitivity and specificity (-ST: .88 and .86; -HF: .81 and .87), respectively (Vaughan, 2011). Authors reported finding a two-factor solution for the CARS-ST and a three-factor solution for the CARS2-HF (Schopler et al., 2010).

Beneficial uses of this scale include the ability to use information from multiple raters, good reliability and validity, short and easy administration, and the two different versions account for age, cognition, and language (Schopler et al., 2010). Limitations include that the normative sample consists of mostly White males, the lack of structured activities for the clinician to observe the child (e.g., as the ADOS-2 includes), and that raters need to know specific information before choosing a form – potentially introducing difficulty and/or additional assessment (Malcolm, 2014; McClellan, 2014; Schopler et al., 2010).

GARS-3. The GARS-3 is a rating scale designed for parents/caregivers, school staff, or other professionals to rate individuals between the ages of 3 and 22 years. It has 58 items and six different subscales that contribute to an overall score, with the number of subscales completed dependent on the individual's level of communication and language. The GARS-3 is intended to be used to identify likelihood of ASD, but also provide information relevant to diagnosis and determination and/or measurement of education or treatment services and goals (Gilliam, 2013). As the GARS-3 is the focus of this project, more information about the assessment and its psychometric properties will be explored in the section related to the different GARS editions.

Advantages and Disadvantages of Rating Scales. In all, rating scales have a number of utilities and benefits. They are beneficial in gathering information from multiple raters and contexts, and likely provide contextualized information about the individual, more nuanced than a singular observation (Norris & Lecavalier, 2010b; Scahill & Lord, 2004). Rating scales may take less time to administer and are typically less expensive compared to diagnostic interviews and observations, especially when it comes to training examiners (Charman & Gotham, 2013; Kuriakose & Shalev, 2016; Scahill & Lord, 2004). However, some scales lack solid psychometric support, particularly through independent research, regarding normative data, validity, and reliability. Rating scales collect important information, and clinicians should be knowledgeable about the intention of the scale – as a screener, as part of diagnostic assessment, or informing clinical or educational treatment planning. Because of the recognized utility of rating scales for these purposes, independent psychometric evidence is critical.

Diagnosis

When screeners indicate a need for further evaluation, a clinical diagnostic assessment is warranted. Common assessment tools to aid in ASD diagnosis, in addition to rating scales, include interviews and observations (Kuriakose & Shalev, 2016). For best practice in ASD evaluation, interviews *and* observation techniques should be used – as to not decrease specificity by only using one of these methods (Huerta & Lord, 2012). In all, it is important to use information from multiple sources and to gather data using various methods.

Two of the most commonly used ASD diagnostic assessment tools are the Autism Diagnostic Interview-Revised (ADI-R; Rutter et al., 2003b) and the Autism Diagnostic Observation Schedule – Second Edition (ADOS; Lord et al., 2012). These assessments have the most research support and high ratings of validity in estimates of sensitivity and specificity. Used together, they have a highly accurate classification rate (Falkmer et al., 2013).

Interview. In general, interviews are used in assessment to gather information through conversations with stakeholders who know the individual well. Interviews can be structured, semi-structured, or unstructured. In structured interviews, there are preset questions that are asked in a designated order with standardized response options. Semi-structured interviews have preset questions, however, involve open-ended responses. Unstructured interviews consist of questions that are somewhat defined, with open-ended responses meant to be given in a discussion format. Rating scales can often be given in the form of a structured interview and structured interviews generally lead to the most accurate comparison across stakeholder information (Salvia et al., 2017).

ADI-R. The Autism Diagnostic Interview – Revised (ADI-R; Rutter et al., 2003b) is a semi-structured structured interview used in the ASD diagnostic process (Ozonoff et al., 2005). Its original form was developed in 1994 and the revised version’s changes include omitting items, a reorganization of the interview structure, and some wording changes. The algorithm to calculate an overall score was unchanged (Rutter et al., 2003b). Within the interview, a trained clinician talks with the parent or caregiver of the target individual. The long version, used for research, may take up to three hours. The shorter version, which may be used for clinical assessment, includes select items to complete the algorithm score and may take ninety minutes to complete and score (Lord et al., 1994). Clinicians elicit information regarding the developmental history of the individual, in addition to current behavior (e.g., social communication, restricted and repetitive behaviors, language, interests; Ozonoff et al., 2005; Rutter et al., 2003b). After coding each item, the clinician calculates scores used in the algorithm, generating interpretable results for diagnosis or assessment of current behavior (Rutter et al., 2003b).

Originally, the ADI (Lord et al., 1994) was standardized on a sample 20 children (10 with ASD and 10 without ASD). In a follow-up validity study, 30 more individuals (half with an ASD diagnosis) were assessed. Interrater reliability, as reported by Lord et al. (1994), were higher for the domain and subdomain scores, compared to the behavioral items, and intraclass correlations ranged from .93 to .97. In a study by Lord et al. (1993) in a population of preschool-aged children, researchers reported test-retest reliability intraclass correlation coefficients as greater than .92 for the domain and subdomain scores. Examination of discriminant validity provided evidence that most domain and subdomain scores accurately discriminate between ASD and non-ASD groups (including those with developmental disorders). However, there were not statistically significant differences between the groups for individual items related to stereotyped repetitive use of language and reciprocal conversation – although the sample size was small. Overall, the algorithm score was found to highly discriminate between groups (Rutter et al., 2003b).

As a gold-standard measure, the ADI-R has clear benefits. It provides relevant information pertaining to the developmental period that may be indicative of a developmental delay. The overall algorithm score was also found to be valid data to contribute to an ASD diagnosis (Rutter et al., 2003b). However, some limitations include its lack of sensitivity for lower-functioning individuals (e.g., IQ below 20), and requirement of in-depth training (e.g., three days for clinical use). Moreover, it can require a substantive amount of time to administer and score (e.g., 1.5-3 hours; Ozonoff et al., 2005). In all, the ADI-R is a key element in an ASD diagnostic assessment in order to obtain information about developmental and current behavior.

Observation. Observations are an assessment technique that provide information about and contextualize behavior of a target individual. Observations can be defined as systematic or nonsystematic. In a systematic or formal observation, the observer is looking for specific and objectively defined behaviors. Additionally, they observe what precedes or follows the behavior, and measure the behavior’s frequency, duration, amplitude, or latency. In a nonsystematic or informal observation, the goal is to observe the individual in specific environment or situation and take note of behaviors (Salvia et al., 2017).

ADOS-2. The Autism Diagnostic Observation Schedule – Second Edition (ADOS; Lord et al., 2012) is a systematic observation tool used to gather information about behaviors related to ASD in diagnostic assessment. Observation of the assessed individual is imperative to ASD assessment, specifically for the clinician to observe behaviors that define ASD – deficits in social interaction and communication, in addition to restrictive interests and repetitive behavior. ASD is characterized by these behaviors, and as there is no current definite medical or biological test for diagnosis, clinicians rely on gathered information that establishes the presence of the behavior (APA, 2013).

The gold-standard ASD observation measure is the ADOS-2, a semi-structured observation previously published in its first edition in 1999 (ADOS; Lord et al., 1999; McCrimmon & Rostad, 2014; Ozonoff et al., 2005). Additions to the revision include updated algorithms and the new Toddler module (McCrimmon & Rostad, 2014). The ADOS-2 involves structured activities, set-up by the clinician, who then observes and records the individual's responses as well as other behaviors. By observing and eliciting behavior through the activities, the clinician rates several areas related to ASD criteria. The ADOS-2 has four different modules, in addition to the Toddler module, which are selected based on age and language level of the child (McCrimmon & Rostad, 2014; Ozonoff et al., 2005). For Modules 1 and 2 with children of younger mental and/or chronological age, the ADOS assesses behaviors such as communicative behavior, symbolic play, joint attention, social interest, and atypical behaviors. For Modules 3 and 4, the observation is more focused on reciprocal conversation skills, empathy, insight into different roles and relationships, and special interests (Ozonoff et al., 2005).

The ADOS-2 also demonstrates good psychometric reliability and validity. The ADOS-2 validation (*n* = 1,574) and replication (*n* = 1,282) samples included ASD, “nonautism ASD,” and “nonspectrum diagnoses” (McCrimmon & Rostad, 2014, p. 90). The majority of the samples had an ASD diagnosis and were Caucasian males. Authors reported high internal consistency reliability for social and communication domains (e.g., Cronbach α = .87-.92 for modules 1-3), and moderate values for domains associated with restrictive and repetitive behaviors (e.g., Cronbach α = .51-.66 for modules 1-3). Test-retest reliability was reported as moderate to high (e.g., correlations = .68-.92 for domain and overall scores in modules 1-3). Interrater reliability for all five modules met or exceeded .71 at the item-level and ranged from .79-.98 for domain and overall scores for modules 1-3. Regarding validity, exploratory and confirmatory analyses support a two-factor structure. Researchers reported that the overall score holds the highest predictive value and as such, should be used in decision making. Authors also reported predictive validity (e.g., sensitivity = .60-.95 and specificity = .75-1.00 for modules 1-2; McCrimmon & Rostad, 2014).

Benefits of using this standardized observation tool include that it is based on direct observation and independent of potential biases of different reporters. Additionally, the appropriate trained clinician will likely provide reasonably objective data regarding the child’s behavior (Scahill & Lord, 2004). Limitations of this measure are primarily the significant costs and substantial time investment in training and implementation of the measure. Like the ADI-R, the ADOS-2 requires specific training, which can be costly, and requires expertise and practice to effectively implement and code each module. Beyond the time and monetary cost of training, the cost of the kit and the amount of time involved in administering the observation may also limit its use (Charman & Gotham, 2013; Scahill & Lord, 2004). In all, research suggests that the

ADOS-2 is a reliable and valid measure to be used in clinical and research settings (McCrimmon & Rostad, 2014).

Progress Monitoring and Outcome Assessment

Another type of assessment is monitoring progress of treatment. Progress monitoring can be used to measure growth of a student's goals and further, to make decisions about what skills the student needs to acquire and the appropriate level in which to intervene. In an educational setting, progress monitoring may be used to measure these treatment goals but might also inform educators about a student's progress related to educational standards (e.g., common core or state standards; Salvia et al., 2017).

In the evaluation of a specific intervention or treatment, clinicians and researchers can use a variety of different assessments to measure outcome variables. These outcome measures should treatment sensitive, and in the best-case circumstance, rated by a clinician. The type of the outcome measure will likely depend on the focus of the intervention but with students with ASD, may measure core symptomatology or associated features (Scahill & Lord, 2004; Smith et al., 2007). Further, an ideal assessment would accurately measure outcomes in treatment *and* in skill generalization within an individual's typical environments (Lord et al., 2005; White et al., 2007). However, clinician rated measures may not be feasible as they are timely in their training and costly for an expert clinician to administer. Measures used for outcome assessment may also lack data to support their sensitivity to change (e.g., ADOS) or reliability and validity (e.g., Clinical Global Impression for Severity (CGI-S) and Improvement (CGI-I); Guy, 1976; Scahill & Lord, 2004). Other possibilities for outcome assessment measures include coded observations, ratings from multiple informants, and the use of self-report data (White et al., 2007).

Comprehensive ASD Assessment: Associated Domains and Comorbid Features

In comprehensive ASD assessment, there are domains of associated and comorbid features that should be assessed, beyond those related to core features of ASD. These often include adaptive functioning, cognitive functioning, developmental abilities, language, and verbal and nonverbal abilities (Huerta & Lord, 2012; Kuriakose & Shalev, 2016). Further assessment may also be warranted for commonly occurring comorbidities (e.g., ADHD, anxiety; Zwaigenbaum & Penner, 2018).

Developmental assessment is important to gauge an individual's level of functioning, provide context for behavior, and provide information for treatment (Kuriakose & Shalev, 2016; Ozonoff et al., 2005). Current ASD diagnostic specifiers (e.g., presence of cognitive impairment), relate to these areas and as such, should be assessed (APA, 2013; Zwaigenbaum & Penner, 2018). Moreover, these characteristics are important for informing outcomes as cognition is shown to be correlated with symptom severity and the ability to learn skills (Ozonoff et al., 2005). Two common assessment tools to measure development and cognition, which are also ideal for those with suspected ASD because of lower language demands, include the Mullen Scales of Early Learning (MSEL; Mullen, 1995) and the Differential Ability Scales, Second Edition (DAS- II; Elliott, 2007) which also examines academic performance (Kuriakose & Shalev, 2016; Ozonoff et al., 2005).

Additionally, adaptive functioning is important to assess as it is important outcome data and imperative for treatment planning and skill development (Ozonoff et al., 2005). For adaptive skills assessment, a very common assessment tool is the Vineland Adaptive Behavior Scales, currently in its third edition (Vineland-III; Sparrow et al., 2016), which includes forms for teachers and parents to complete interviews or rating scales. It can provide information regarding communication, daily living skills, social skills, and fine and gross motor skills (Kuriakose & Shalev, 2016; Ozonoff et al., 2005).

Language is also important to assess as it is a key outcome predictor and a specifier in ASD diagnosis (APA, 2013; Ozonoff et al., 2005). In the assessment of language within the context of an ASD diagnosis, clinicians may use different assessment tools, but might also refer to a specialist to complete more in-depth assessment and provide recommendations. Common tools to assess receptive and expressive language, listed here in their current editions, include the Peabody Picture Vocabulary Test (PPVT-5; Dunn, 2018), Expressive One-Word Picture Vocabulary Test (EOWPVT-4; Martin & Brownell, 2010), Clinical Evaluation of Language Fundamentals (CELF-5; Wiig et al., 2013), and Preschool Language Scales (PLS-5; Zimmerman et al., 2011; Ozonoff et al., 2005).

Overall, there are different procedures and domains of assessment within ASD evaluations. Rating scales, interviews, observations, and standardized assessment batteries are used to provide information about core (i.e., social communication and interactions, restrictive and repetitive behaviors) and associated features (e.g., cognition, language, adaptive skills).

Gilliam Autism Rating Scale (GARS)

The assessment measure of focus for this series of studies is the Gilliam Autism Rating Scale (GARS). Characterized as a level two screener, the GARS can be used as part of the screening and diagnostic process (Gilliam, 2013). Newer editions of the GARS have been cited among the most commonly used ASD rating scales in school or clinical settings (e.g., Aiello et al., 2017; Benson et al., 2019; Kuriakose & Shalev, 2016). Prior editions of the GARS have also been translated and used in other countries (e.g., Diken et al., 2012; Jackson et al., 2013). The manual also suggests utility of the GARS in annual evaluations and/or educational planning – although there is little independent research to support its sensitivity for this use (Gilliam, 2013).

Although the newest version of the GARS, the GARS-3 (Gilliam, 2013), has been published since 2013, there is little to no published research regarding its psychometric properties. Although it is marketed for use in assessment and intervention, practitioners only have reliability and validity estimates based on the manual and its standardization sample. Thus, it is important to contribute to the paucity of psychometric research for the GARS-3 in order to provide needed independent support for its use and value in clinical and research settings. The following sections illustrate changes between editions and provide an in-depth analysis of the psychometric properties of the current edition as reported in the manual.

GARS

The original GARS, published by James Gilliam in 1995, was intended for individuals three to 22 years of age and consisted of four subscales. The Stereotyped Behaviors, Communication, and Social Interactions subscales aligned with the three core symptom clusters of autistic disorder in DSM-IV (i.e., impairments in social interactions and communication, and the presence of restricted repetitive and stereotyped patterns of behavior). The Developmental Disturbance subscale assessed early developmental history and was intended to identify delayed or abnormal development (Lecavalier, 2005; South et al., 2002). Raters of the scale answered 56 items on a 4-point rating scale that ranged from 0 (never observed) to 3 (frequently observed; South et al., 2002). Subscales were calculated by summing and converting scores into standard scores with a mean of 10 and standard deviation of 3. The total score or Autism Quotient (AQ) has a mean of 100 and standard deviation of 15 and reported to be “reliably and validly calculated” from using a combination of two, three, or all four subscales (South et al., 2002, p.

595). The selection of subscales was dependent on the language ability of the child (e.g., Communication subscale might not be appropriate if the child was nonverbal) or if the rater was not familiar with developmental history (e.g., rating the Developmental Disturbances scale would not be appropriate). The AQ was indicative of the “likelihood” of autism with a cut score of at least 90 indicating the individual is “probably autistic.” The AQ also had ordinal, qualitative categories from “Very Low” to “Very High” (South et al., 2002).

The standardization sample of the GARS consisted of 1,092 individuals living in the US or Canada, whom had parent- or teacher-reported autism diagnoses (South et al., 2002). Reliability for the scale was reported as high; coefficient alpha ranged from .88 to .93 and test-retest reliability was reported as .81 for the behavioral subscales and .88 for the total score. Inter-rater reliability, with a small sample, was reported as ranging between .73 and .82 for the behavioral subscales and was .88 for the total score. Validity for the scale was also reported. The manual provided evidence to support convergent validity with the ABC (e.g., total score correlation of .94), as well as discriminant validity (e.g., differentiating autism from other disorders) and construct validity (e.g., they reported no correlation between age and subscale or total scores; Lecavalier, 2005).

According to the manual, the GARS was recommended by most state education agencies and frequently used in research (Gilliam, 2013). For example, the second edition manual indicated “generally positive” reviews. Specifically, reviews praised it as a nice addition to autism assessment instruments and that its development and psychometrics were adequate. Further, reviews noted its efficient and flexible use in addition to its contribution to the lack of existing ASD assessment instruments (Gilliam, 2006).

However, there were also several research studies which suggested limitations of the instrument. There was criticism surrounding the methods of the data presented in the manual including small sample sizes, use of standard scores in reliability calculations, and means of reporting interrater reliability (Lecavalier, 2005). Additional researchers (e.g., South et al., 2002) noted concern regarding the "probability of autism" rating (Gilliam, 2006). Their research suggested that Gilliam's cut score guidelines for the subscale standard scores and AQ led to under-identification of cases of autism (i.e., high rates of false negatives; Gilliam, 2006; South et al., 2002).

Lecavalier (2005) specifically examined construct and diagnostic validity. Through factor analysis, results indicated a different factor structure than reported in the manual and further, found a large number of items loading onto the factor representative of stereotyped and repetitive behavior. Additionally, there was criticism regarding the contribution of the Developmental Disturbance subscale – it seemed to not make significant contributions to the total score and had low internal consistency. The average AQ from Lecavalier's (2005) sample was lower than what was reported in the manual which may provide additional evidence for low sensitivity estimates. More information on Lecavalier's factor analytic study will be provided in a later section.

GARS-2

The second edition of the instrument, the GARS-2, was published in 2006. The manual cited several differences in this new version, noting considerations for independent findings, reviews, observations, and questions. Revisions included replacing the Developmental Disturbances subscale with a developmental parent interview (that no longer contributed to the composite score), revision of select items, new norms based on the 2000 U.S. Census, a new label of the total score (i.e., Autism Index), new guidelines for subscale and index scores, specific definitions and examples of behaviors referenced in the items, and a booklet intended to assist with developing instructional goals and objectives (Gilliam, 2006). Further, this revision included optional separate scoring and reporting software. The test author reported that the GARS-2 was "accepted" by parents and professionals, used in different countries including non-English speaking countries, and that the addition of the booklet to help develop instructional goals and objectives helped link assessment and intervention services (Gilliam, 2013).

The GARS-2 consisted of items 42 items and three subscales (i.e., the same behavioral subscales as the previous edition – Stereotyped Behaviors, Communication, and Social Interaction). As mentioned, there was a parent interview to provide historical developmental information which did not contribute to the total score but provided information about and highlighted specific difficulties. As in the previous edition, subscale items were summed to create scaled scores, which were then summed to create a total score (i.e., the Autism Index). If the Communication subscale was deemed inappropriate because of the individual's verbal abilities, the two remaining subscale scores could be summed to produce the Autism Index (Gilliam, 2006). The cut score associated as the highest probability (i.e., "Very Likely") of autism was an Autism Index of 85 or higher – compared to the score of 90 from the previous edition. This change in cut score may have been influenced by independent researchers' findings and critiques of the GARS's under-identification of individuals with ASD (e.g., Lecavalier, 2005; South et al., 2002; Gilliam, 2006).

The GARS-2, similar to the GARS, relied on diagnostic definitions of autism from the Autism Society of America in addition to the updated DSM-IV-TR. The standardization sample consisted of 1,107 individuals from the United States between the ages of three and 22 (Gilliam, 2006). Ratings were gathered by posting on an Asperger's Syndrome-related website for parents, by professionals contacting the test author, and by contacting teachers and other professionals (e.g., psychologists, educational diagnosticians – 42 of whom submitted data for 692 individuals; Gilliam, 2006; Montgomery, 2008). Besides meeting the age inclusion criteria, individuals needed to have an autism diagnosis, and live in the US (Gilliam, 2006). No information was reported regarding language delays of the individuals and there is a small percentage (i.e., 9%) representative of older individuals ages 16 through 22 (Montgomery, 2008).

The manual also reported reliability and validity of the instrument. Internal consistency reliability yielded Cronbach alpha coefficients of .84 - .88 for subscales and .94 for the total score. Additionally, test-retest reliability was reported from a sample of parents of 37 individuals, two weeks apart. The Pearson correlation coefficient for the Autism Index was .84, while the subscales ranged from .64 - .83. No inter-rater reliability was reported. Content validity was demonstrated by the items' alignment with popular definitions of autism. Criterion-related validity was established with concurrent correlations with the ABC. Construct validity was demonstrated by showing low correlations between GARS-2 scores and age (i.e., -.13 - .06), high correlations between subscales (i.e., .46 - .56), and high correlations between the items and the subscale scores (i.e., median item-discrimination coefficients, .53 - .55). Further, the manual reported that the GARS-2 was able to discriminate between individuals with autism and those who were non-disabled, multi-disabled (e.g., blind, deaf, internalizing and externalizing disorders), and "mentally retarded" – as evidenced by significantly higher subscale and total scores. Positive predictive outcome analyses were used to examine how the GARS-2 was able to correctly identify individuals with autism in samples without autism (i.e., non-disabled, mentally retarded, multi-disabled), which yielded the sensitivity levels ranging from .84 – 1.00, specificity levels ranging from .84 - .87, positive predictive values from .84 - .85, and percentage of diagnostic agreement from 84 – 93%. No factor analytic evidence of construct validity was provided for this sample (Gilliam, 2006).

Several studies independently examined the psychometric properties of the GARS-2. Pandolfi, Magyar, and Dill (2010) examined the GARS-2 factor structure using exploratory and confirmatory factor analysis. Their results supported four factors which did not align with the three-factor solution published in the manual. Pandolfi et al. (2010) also questioned whether the instrument could assess students considered higher functioning. Additionally, researchers were critical of the use of "double-barreled" items wherein the item referenced two different behaviors. In this case, the rater might be attending to one behavior more than the other, potentially leading to discrepancies between different raters. Moreover, Pandolfi et al. (2010) were critical of item placement on subscales such that some items seemed to be related to, but not a core feature of a specific construct. These issues may have adversely impacted GARS-2 discrimination between those with and without ASD (Pandolfi et al., 2010). Likewise, in two factor analytic studies by Volker et al. (2016 and 2020), researchers found discrepancies regarding item placement on factors compared to the published GARS-2 factor structure. They also reported high internal consistency reliability and found that a three-factor solution best fit the data. Volker et al. (2016) reported finding a higher sensitivity estimate than reported in the manual, however, also reported high numbers of false negative cases –34.71% of the ASD sample were inaccurately classified as non-ASD. More information about factor analyses of the GARS-2 will be reported in a later section.

GARS-3

The GARS-3, published in 2013, is the most recent edition of the instrument. Like previous versions, it is characterized as a norm-referenced screening measure used to identify the likelihood of individuals having ASD (Gilliam, 2013). This edition reflected an updated conceptualization of ASD and was informed by DSM-5 criteria, which shifted multiple diagnoses related to autism (e.g., Asperger's, autistic disorder) to autism spectrum disorder. Consistent with prior editions, the GARS-3 is used with individuals ages three through 22 years, with individuals who demonstrate severe behavioral problems potentially indicative of autism, and with parent/caregiver or teacher raters who have had regular contact with the individual for at least two weeks. Additionally, items are similarly scored on a 4-point rating scale from 0 (not at all like the individual) to 3 (very much like the individual). With this new edition, Gilliam indicated that attempts were made to address validity concerns conveyed in reviews and research findings pertaining to the GARS-2, as indicated in the above section (Gilliam, 2013).

There were several major changes with the GARS-3. The GARS-2 consisted of 42 items distributed across three subscales—which reflected the three core symptom clusters of autistic disorder, as conceptualized under DSM-IV (Gilliam, 2006). In contrast, the GARS-3 contains 58 items – 16 of which were kept from the GARS-3 and 42 newly added. Instead of three subscales, the GARS-3 is divided into six subscales with items that "describe specific, observable, and measurable behaviors" (Gilliam, 2013, p. 2). New subscales reflect social communication, emotional responding, cognition, and speech patterns – added in addition to the previously measured subscales of restrictive and repetitive behaviors and social interaction. Constructs measured by the new subscales were reportedly added because of empirical evidence of their validity and sensitivity in ASD identification (Gilliam, 2013). The manual explains that while emotional responses are not part of an ASD diagnosis, they strongly contribute to "the overall diagnostic picture" (Gilliam, 2013, p. vii). Additionally, the Stereotyped Behaviors subscale was changed to Restricted/Repetitive Behaviors to better align with DSM-5 language. In this new version, Gilliam incorporated new normative data, used analyses and studies to improve validity (e.g., exploratory factor analysis [EFA], binary analysis), and aesthetically, used a new updated look (Gilliam, 2013).

Each of the six subscales is intended to measure a construct related to ASD. The first subscale, Restricted/Repetitive Behaviors contains 13 items and reflects stereotyped behaviors, restricted interests, routines, or rituals (e.g., staring at hands or other things for at least 5 seconds, flicks fingers in front of eyes, unusual sensory interests). The second subscale, Social Interaction consists of 14 items and reflects social behavior (e.g., initiating conversations, expressing pleasure in interactions, showing interest in others). The third subscale is Social Communication, containing 9 items, and reflects social responses, understanding, and communication (e.g., understanding jokes or teasing, prediction of what might happen in social situations). The fourth subscale, Emotional Responses, contains 8 items and reflects extreme and everyday emotional responses (e.g., excessive reassurance for change, extreme reactions to loud, unexpected noise). The fifth subscale, containing 7 items, is Cognitive Style which relates to “idiosyncratic fixed interests, characteristics, and cognitive abilities” (e.g., precise speech, restricted interests reflected in conversation). The sixth and last subscale is Maladaptive Speech which consists of 7 items and reflects “deficits and idiosyncrasies in verbal communication” (e.g., echolalia, flat tone or affect, use of idiosyncratic words; Gilliam, 2013, p. 3).

Examiners can use these six subscales to score and interpret ratings. The GARS-3 composite score, the Autism Index, can be calculated using four (Autism Index 4) or six (Autism Index 6) subscales. This is dependent on the level of communication of the rated individual. If the individual is not communicative, then the rater only completes the first four subscales and uses the Autism Index 4 as the total score (Gilliam, 2013). It is estimated that 40% of individuals with ASD are nonverbal or “mute” and that 25-30% of children with ASD have some words at 12 to 18 months, but speech is halted at some point in development. For others, speech might not occur until later in development (CDC, 2019; Gilliam, 2013). In the normed data used in the GARS-3, 25% of the sample was characterized as mute and many did not sign or use augmentative communication devices to communicate. The Maladaptive Speech subscale and three Cognitive Style subscale items (i.e., 45, 47, 51) require some level of communication and are therefore not appropriate to rate. The manual indicated omission of these subscales for someone who is “mute” and who has not communicated through signs or speech in the past six months (Gilliam, 2013).

Overall, the Autism Index is the most reliable standard score and both versions (i.e., Autism Index 4, Autism Index 6) were described as “reliable, valid, and discriminative” (Gilliam, 2013, p. 15). The scores are standardized and calculated in the same manner as previous versions (i.e., scaled scores with a mean of 10 and *SD* of 3; standard scores (Autism Index) mean of 100 and *SD* of 15). These scores are reported to be useful in comparisons, instead of percentile ranks. Additionally, scaled scores can be beneficial for use in research (Gilliam, 2013).

With the Autism Index score, and using the interpretation guide, the examiner can determine – based on predetermined cut scores – how likely it is for the individual to have ASD, the severity level (i.e., associated with diagnostic criteria in the DSM-5 for ASD), and a descriptor for how much support is needed. When using the Autism Index, the interpretation of the score is based on a normative sample of individuals with ASD. As such, the higher the Autism Index score, the greater the likelihood of ASD. Based on the Autism Index score, an individual’s likelihood of ASD can be classified as unlikely (e.g., score below 55), probable (e.g., score 55-70), and very likely (e.g., score greater than or equal to 71). With an Autism Index score below 55, there is still possibility for a diagnosis, however, it is unlikely. According to Gilliam, only four cases in the ASD normative sample had scores below 55 – which the author suggests may have occurred due to misdiagnosis or substantial improvements in the rated behaviors since the original diagnosis. With a "probable" Autism Index score of 55-70 – 2-3 *SD* from the mean – there might be a milder level of severity as individuals may appear high-functioning or more similar to their typically developing peers. The "very likely" category encompasses Autism Index scores greater than 70 with two different descriptors: requiring substantial or very substantial support. An Autism Index score 71 through 100 – with scores is 1-2 *SD* from the mean – is indicative of requiring substantial support and subcategories are conceptualized according to individual presentation. Scores between 71 and 84 characterize individuals who may be considered "high functioning" – they have intact language and typically do well in school, but likely have the most difficulty with social communication and emotional responding. Gilliam notes that before the DSM-5, this range would encapsulate those having diagnoses such as Asperger's, Rett's, or PDD-NOS. Scores 85 through 100 indicate significant behavior associated with autism with limited academic and social interactions. Scores greater than or equal to 101 are also considered "very likely" to have ASD but with greater severity of behavior. Gilliam describes individuals with scores in this category to require significant support, attention, and programming with "little doubt about the diagnosis" (Gilliam, 2013, pp. 16-17).

The manual reported several uses for the GARS-3 including assisting with the identification of ASD, assessing level of severity, demonstrating intervention progress, specifying IEP goals, and for use in research. Gilliam first and foremost describes the GARS-3 as a way to provide objective data to identify likelihood of ASD, used in context with other data.

This measure also assesses severity level, which is useful to provide recommendations for treatment, especially when looking at specific subscales and items. This data can also be used in an educational setting to plan a student's program and for yearly evaluation. More specifically in an educational setting, the GARS-3 could be used to identify strengths and weakness, set IEP goals, and identify the targets of intervention. The manual indicated support for its use in research as the GARS-3 is reliable and valid. The instrument could be especially useful since behaviors are measured in frequency (Gilliam, 2013). More information about how the GARS-3 has been utilized in research will be provided in a subsequent section.

Psychometric Properties. Norming procedures, evidence of reliability and validity, and related studies were reported in the GARS-3 manual. The normative sample included 1,859 children and young adults with ASD. The inclusion criteria consisted of individuals with a diagnosis of autism, between the ages of three and 22 years, and residing in the United States. The sample was recruited through announcements on websites and by email. Email lists were compiled of subscribers to ASD-related journals and current users of the GARS-2 as noted in PRO-ED customer files. Data were collected from 1,859 parents (1.9% fathers, 21.3% mothers), teachers (33.4%), educational diagnosticians (1.5%), psychologists (1.8%), speech clinicians (13%), teacher assistants (8.7%), other school and treatment center personnel for students with ASD (18.5%). The majority of raters had advanced degrees (58.6%), reportedly were "very knowledgeable" about ASD (71.4%) and had six or more years of experience with individuals with ASD (58.9%). Data was collected both online (93%) or through a paper and pencil form (7%; Gilliam, 2013).

Within the sample, 61.3% had a sole diagnosis of ASD and 38.75% had an ASD diagnosis and one or more other diagnoses (Gilliam, 2013). No other information was provided about specific comorbid disorders of the sample (Gilliam, 2013; Hutchins, 2017). The manual noted that the decision to include both of these groups in one larger sample was because of the increase in comorbidity prevalence rates. The number of individuals in each age range was also reported. Most age groups consisted of at least 85 individuals (i.e., ages 3-18). Groups from 19 – 22 years of age ranged from 25 to 44 individuals. Although this is important to note, scores are not normed for different ages. It was reported that the normative sample represented the larger ASD population regarding a wide range of symptoms, the geographic area, cultures, race, and gender (Gilliam, 2013). The gender distribution of the sample is 77% male and 23% female. Although not aligned with the Census data of about equal distribution of males to females, this aligns with the data from the CDC – at the time, the reported ratio was 5:1 and current CDC prevalence estimates indicate a 4:1 ratio of males to females (CDC, 2020; Gilliam, 2013). Other demographic characteristics closely resembled Census data for geographic region, race (i.e., predominantly White (80%)), and Hispanic origin (Gilliam, 2013).

Reliability. Estimates of internal consistency (coefficient alpha), test-retest, and interrater reliability for the GARS-3 were reported in the test manual. Alpha coefficients were calculated individually for ages 3-18 and grouped for ages 19-22 using data from the normative sample. The manual reported coefficient alphas for subscales and Autism Index scores, and additionally reported an average alpha value for the sample using Fisher's *z* transformation. Subscale alpha coefficients ranged from .88 - .93 ($\bar{X} = .90$) for Restricted/Repetitive Behaviors, .91 - .96 ($\bar{X} = .94$) for Social Interaction, .86 - .92 ($\bar{X} = .89$) for Social Communication, .86 - .94 ($\bar{X} = .90$) for Emotional Responses, .80 - .89 ($\bar{X} = .86$) for Cognitive Style, and .71 - .85 ($\bar{X} = .79$) for Maladaptive Speech. Total score alpha coefficients ranged from .93 - .96 ($\bar{X} = .94$) for the Autism Index 4 and .90 - .95 ($\bar{X} = .93$) for the Autism Index 6 (Gilliam, 2013).

For test-retest reliability, data were collected for 122 individuals. Raters completed two scales with 1-2 weeks between ratings. Test-retest raters consisted of 44 parents, 48 teachers, 4 speech clinicians, 9 TAs, and 17 other professionals (e.g., consultants, therapists, supervisors). Ages of the rated individuals were three to 22 years of age, with a mean age of 11.9 (*SD* = 5.4). The sample included a majority of cases with an ASD diagnosis (*n* = 114) and also included other groups (*n* = 3 ID, *n* = 2 ADHD, *n* = 2 TD, and *n* = 1 gifted and talented). Results for correlations were reported both as uncorrected correlations (r_u), and corrected correlations (r_c) which were corrected for range and attenuation. Subscale and composite reliability correlations, uncorrected and corrected, are as follows: Restricted/Repetitive Behaviors ($r_u = .85, r_c = .83$), Social Interaction ($r_u = .81, r_c = .78$), Social Communication ($r_u = .83, r_c = .87$), Emotional Responses ($r_u = .83, r_c = .76$), Cognitive Style ($r_u = .80, r_c = .84$), Maladaptive Speech ($r_u = .76, r_c = .80$), Autism Index 4 ($r_u = .89, r_c = .90$), and Autism Index 6 ($r_u = .91, r_c = .90$; Gilliam, 2013). As indicated by Salvia et al. (2017), results demonstrated good reliability for composite scores and most subscale scores, as correlations should be at least .70 for weekly progress monitoring, .80 for screening, and .90 for important educational decisions. These scores also demonstrate good reliability for use in research, according to Nunnally's (1978) standard.

The manual included an interrater reliability study, which involved a sample of 116 individuals with ASD and 232 raters from 23 states (e.g., parents, teachers, psychologists, speech clinicians, teacher assistants, others). Rater pair combinations did not solely occur within the same subgroup (e.g., pairs included parent-parent, teacher-teacher assistant, teacher-psychologist, etc.). The interrater ASD sample had a mean age of 11.4 (*SD* = 4) and was 80% male and 88% White. The interrater intraclass correlation coefficients (ICC) for the subscales and composite scores were reported as follows: Restricted/Repetitive Behaviors (ICC = .84), Social Interaction (ICC = .75), Social Communication (ICC = .85), Emotional Responses (ICC = .85), Cognitive Style (ICC = .83), Maladaptive Speech (ICC = .71), Autism Index 4 (ICC = .84), and Autism Index 6 (ICC = .84; Gilliam, 2013). Comparing these values to the reliability standards as reported by Salvia et al. (2017), Autism Index composite scores, along with some subscale scores, met criteria for reliability used for weekly progress monitoring and/or screening (i.e., above .70 for progress monitoring, above .80 for screening). Further, scores meet Nunnally's (1978) standard for use in research. However, no values exceeded the .90 educational decision standard (Gilliam, 2013; Salvia et al., 2017).

Validity. The manual also included evidence for the validity of the GARS-3. Specifically, validity was reported for item content, criterion-related, and construct-identification (Gilliam, 2013). Item content, or content validity, refers to the examination of the scale's constructs. Is the scale measuring what it intends? Is this reflected in the items? Is there any content missing? How are these constructs measured (Salvia et al., 2017)? In scale development, the author examined ASD domains, the selection and appropriateness of items, and item analysis. The content of the GARS-3 covers ASD domains of social communication and interaction in addition to restrictive and repetitive behaviors, aligned with the DSM-IV-TR and DSM-5. In the selection of specific items, a checklist of 120 items was developed based on diagnostic criteria and previous diagnostic tests. Tests included the CARS-2 (Schopler et al., 2010), the Autism Behavior Checklist from the Autism Screening Instrument for Educational Planning – Third Edition (ASIEP-3; Krug et al., 2008), ADI-R (Le Couteur et al., 2003), ADOS (Lord et al., 1999), the Asperger Syndrome Diagnostic Scale (ASDS; Myles et al., 2001), the Gilliam Asperger's Disorder Scale (GADS; Gilliam, 2001), and Krug Asperger's Disorder Index (KADI; Krug & Arick, 2003). This list was sent to clinicians, researchers, and parents for them to indicate their importance. Narrowing the list down to 99 items, the tool was factor analyzed (*n* = 1,516) which produced a six-factor solution with 58 items – with reportedly "sufficiently large" coefficient alpha values (Gilliam, 2013, p. 35). The item selection process is of particular interest as the addition of 42 items was a major change in the third edition. The test author also conducted a conventional item analysis using a point-biserial correlation technique. All items met the conservative cutoff, selected by researchers, of .40 and had a median of .75 (range: .57 - .86; Gilliam, 2013).

Most notable to this dissertation, the GARS-3 manual provided information about the exploratory factor analysis to support construct validity (Gilliam, 2013). The internal structure of a scale should be supported by evidence of hypothesized factors or composites (Salvia et al., 2017). With a sample of 1,859 individuals with ASD (i.e., the normative sample), the data demonstrated a good fit for EFA such that the majority of correlation coefficients in the correlation matrix were .3 and greater, in addition to indicators such as the Bartlett Test of Sphericity and the Kaiser-Meyer Olkin criterion. Results, using maximum likelihood and a promax rotation, indicated six factors, represented by the current six subscales of the GARS-3. The manual indicated that the six factors were in alignment with DSM-5 ASD domains and reported that these findings provide strong evidence for this measure (Gilliam, 2013). These factors do appear to align with the two domains of ASD currently conceptualized in the DSM-5, but additionally, factors seem to also measure associated features and behaviors. A more in-depth analysis of the GARS-3 EFA, contextualizing best practice methods, will be reported in a later section.

Criterion-related validity was examined by looking at correlations between the GARS-3 and other, similar measures (Gilliam, 2013). Criterion-related validity examines the relationship between one assessment and criterion measures. A correlation between the two measures, especially when the criterion measure is well-known, would provide evidence to support that the assessment is measuring the intended construct (Gilliam, 2013; Salvia et al., 2017). The manual indicated the GARS-3 was compared to four criterion measures that yielded large to very large correlations. The following correlations are reported as uncorrected (r_u), followed by the corrected (r_c) correlation. The ASIEP-3 (Krug et al., 2008) yielded very large correlations for the Autism Index 4 ($r_u = .85, r_c = .76$) and Autism Index 6 ($r_u = .86, r_c = .77$). The ADOS (Lord et al., 1999) yielded large correlations with the Autism Index 4 ($r_u = .64, r_c = .72$) and Autism Index 6 ($r_u = .61, r_c = .69$). The CARS-2 (Schopler et al., 2010) also yielded large correlations with the Autism Index 4 ($r_u = .81, r_c = .83$) and Autism Index 6 ($r_u = .66, r_c = .68$). Lastly, the GADS (Gilliam, 2001) yielded very large correlations with the Autism Index 4 ($r_u = .70, r_c = .73$) and Autism Index 6 ($r_u = .75, r_c = .72$; Gilliam, 2013).

The author reported sensitivity and specificity of the GARS-3 through binary classification and receiver operating characteristic/area under curve (ROC/AUC) analyses. Researchers used a cut score of 70 to discriminate between ASD from those without ASD. Results, compared to TD individuals, indicated sensitivity – the ability to accurately identify ASD – of .96 and .95 for the Autism Index 4 and Autism Index 6, respectively. Additionally, specificity – the ability to accurately identify individuals without ASD – of the Autism Index 4 was reported as .95 and the Autism Index 6 was reported as .97. Researchers also looked at sensitivity and specificity for different diagnostic groups compared to those diagnosed with ASD. These values are reported, as the Autism Index 4 value/Autism Index 6 sensitivity and specificity values, respectively: .96/.96 and .88/.75 for the ADHD group comparison, .96/.96 and .62/.79 for the emotional disturbances/behavioral disorders (ED/BD) group comparison, .84/.83

and .91/.93 for the learning disabilities (LD) group comparison, .96/.96 and .78/.80 for the speech-language impairment (SLI) group comparison, and .96/.96 and .78/.84 for the non-ASD disabled group comparison which included diagnoses such as ID, deaf, blind, ADHD, ED/BD, LD, and physical/health impairment (Gilliam, 2013). Of relevance to this study, the non-ASD disabled group included diagnoses of ID (in addition to other diagnoses as noted). Of the 2,240 individuals in this group, 1,779 individuals using the Autism Index 4 and 1,786 individuals using the Autism Index 6 were true positives (i.e., accurately designated as "at risk") and 296 individuals using the Autism Index 4 and 73 individuals using the Autism Index 6 were true negatives (i.e., accurately identified as "not at risk"). Of those not accurately classified, 68 and 73 individuals using the Autism Index 4 and Autism Index 6, respectively, were designated as false negatives, while 85 and 62 individuals using the Autism Index 4 and Autism Index 6 were considered false positives. Using ROC/AUC analyses to examine the accuracy of prediction, the manual reported values of .82 comparing the ASD sample to typically functioning individuals. The manual indicated that above .80 was considered a good rating according to Compton et al. (2006; Gilliam, 2013). In these analyses, it appears that the sensitivity of the test is typically higher than the specificity – indicating that the GARS-3 might be better at identifying those with ASD versus identifying individuals, who may have other diagnoses, as not having ASD.

Further, the manual reported mean difference comparisons between the ASD normative sample and with ID and other diagnostic groups (i.e., ADHD, ED/BD, LD, SLI, typically functioning). Autism Index scores were significantly higher for the ASD group ($p < .01$). Additionally, scores for the ID group were significantly higher compared to the other diagnostic groups ($p < .01$). Although there were significant differences between the ASD and other groups, the ID sample was the only group that mean scores were greater than the cut score of 70 (i.e., Autism Index 4 score of 89, *SD* = 20; Autism Index 6 score of 87, *SD* = 22; Gilliam, 2013).

Current Use of the GARS-3. In the literature, the GARS-3 has been used predominantly to measure ASD symptomatology and/or severity, however, there are no currently available studies that examine the construct validity of the GARS-3. Several studies reported using the GARS-3 to verify diagnosis, measure ASD severity, assess ASD symptomatology – either to characterize individuals in the sample or to divide the sample into groups based on Autism Index score or severity rating (e.g., Alsaedi et al., 2020; Breeman et al., 2020; Brener et al., 2020; Campanaro et al., 2020; Carlile at el., 2018; Cubicciotti et al., 2019; Dass et al., 2018; Ezzeddine et al., 2019; Kay et al., 2020; Knowland et al., 2019; Lordo et al., 2017; Northgrave et al., 2019; Pfeiffer et al., 2019a; Pfeiffer et al., 2019b; Pfeiffer et al., 2018; Rispoli et al., 2018; Rossi et al., 2017; Torres et al., 2018; Vukićević et al., 2019). Additionally, researchers have utilized the GARS-3 to validate another ASD measure (e.g., Eskow et al., 2019) and as a dependent measure of treatment change (e.g., Duffy et al., 2017; Lieneman et al., 2018; Lordo et al., 2017). Some studies did not show significant treatment changes (e.g., Duffy et al., 2017; Lieneman et al., 2018), however, in a study by Lordo et al. (2017), the subscale, Emotional Responses, was significantly different post-intervention.

As mentioned, there is a paucity in the literature regarding psychometrics of the GARS-3. From the studies listed above, there was little to no information about the psychometrics of the GARS-3 as used in their study. One study by Vukićević and researchers (2019) did report good reliability as indicated by Cronbach's α = .951. Regarding psychometrics properties of the GARS-3, there is one published paper with parent ratings, however, only the abstract is available in English. The abstract indicated than in an Iranian sample of 200 individuals with ASD, the

GARS-3 demonstrated very high internal consistency, and “confirmed the content, convergent, and construct validity” (Minaei & Nazeri, 2018). There is also a poster presentation by Hastings and Campbell (2016) that examined the validity of the GARS-3. With a sample of 20 individuals, 12 with ASD and 8 with non-ASD diagnoses, who were rated by caregivers, researchers found that even though the CARS-2 and ADOS-2 scores discriminated between those with and without ASD, the GARS-3 scores between groups did not significantly differ. Moreover, they found that the GARS-3 had weak correlations with the CARS-2 and ADOS-2. The authors acknowledge further work should be done with larger samples in different settings (Hastings & Campbell, 2016). In all, there is a critical need to add to the independent psychometric support of the GARS-3 and of particular importance to this project, support for its construct validity and factor structure.

Factor Analysis in Scale Development and Validation

Exploratory Factor Analysis

Exploratory factor analysis (EFA) is a statistical analysis used for data reduction and to examine underlying constructs (Floyd & Widaman, 1995; Osborne & Banjanovic, 2016). It is commonly used as a first step in scale development to examine psychometric properties (Osborne & Banjanovic, 2016; Yong & Pearce, 2013). This includes looking at the scale’s structure, through examining relationships between individual items/variables, to investigate potential latent constructs/factors derived from the variables (Floyd & Widaman, 1995; Osborne & Banjanovic, 2016). In general, EFA can be used to simplify an analysis or interpretation and highlight patterns of a large number of variables (Gorsuch, 1983; Yong & Pearce, 2013). The general considerations and procedures involved in EFA are cleaning the data, deciding on an extraction method, the method of rotation (if any), and deciding how many factors to retain.

Then, the researcher interprets the solution, if possible, or might re-run the analyses with a different number of retained factors. In the end, the goal is to replicate and/or evaluate the generalizability of the factor structure (Osborne & Banjanovic, 2016).

Like any statistical analysis, there are assumptions for this procedure. As EFA seeks to uncover latent variables, there is the assumption that there *are* latent variables to uncover based on correlations between items (Osborne & Banjanovic, 2016). Further, there should be univariate and multivariate normality with no outliers and a heterogenous sample (Floyd & Widaman, 1995; Yong & Pearce, 2013). There is not a general consensus within the literature regarding a strict sample size requirement, but there seems to be agreement that the larger the sample, the better (e.g., *n* = 300) – with flexibility for smaller samples when there is strong data that yields multiple high factor loadings (e.g., .80 or greater; Costello & Osborne; Yong & Pearce, 2013). There is a general rule that the analysis should have at least a 10:1 ratio of subjects to items – an early criterion, in which Costello and Osborne's 2005 study found the majority of research at the time followed. More traditional rules indicate a ratio of four or five subjects per variable and using a sample of at least 200 subjects (Floyd & Widaman, 1995). In a simulation study by MacCallum et al. (1999), researchers created a table, based on simulations, to indicate how likely it was to discover an accurate factor solution. Researchers can use this table, knowing communalities and indicators in the model and study (i.e., number of factors, items, and participants) to estimate whether they have a large enough sample (MacCallum et al., 1999).

There are also best practices in extraction and rotation procedures. Rotation is used to simplify the model. There are two general types of rotations: orthogonal (e.g., varimax) used for uncorrelated factors and oblique (e.g., promax) used with correlated factors. Costello & Osborne (2005) indicated that the varimax rotation – a specific orthogonal rotation – was the most used rotation. An orthogonal rotation may produce more easily interpreted results, however, especially in the social sciences, there are many variables and constructs that are hypothesized to be correlated. As such, an oblique rotation, which assumes correlated constructs (e.g., promax), will likely produce results more accurate of the population and therefore, is more likely to be replicated (Costello & Osborne, 2005).

Extraction techniques are guidelines for how many factors to calculate and/or interpret. There are several extraction methods including Maximum Likelihood (ML) – best when the data is normally distributed – and Principal Axis Factoring (PAF) – recommended when there is a violation of the normality assumption (Costello & Osborne, 2005; Yong & Pearce, 2013). When deciding the number of factors to retain, there are multiple criteria and/or tests to examine including the Kaiser criterion, a scree plot, Velicer's minimum average partial correlation (MAP) criterion, and parallel analysis (Costello & Osborne, 2005; Osborne & Banjanovic, 2016). Costello and Osborne (2005) reported that most articles in their search used the Kaiser criterion (Kaiser, 1960) which involves retaining factors with eigenvalues greater than one. This is the default of many software programs, however the literature cites this criterion "among the least accurate methods" (Costello & Osborne, 2005, p. 2). The scree plot (Cattell, 1966) is often used by researchers by examining the plot of eigenvalues and factors, looking for the "break or "point of inflection" in the graph, and retaining the number of factors above this point (Yong & Pearce, 2013).

Velicer's MAP (Velicer, 1976) and parallel analysis (Horn, 1965) are other methods not typically found on common software programs (Costello & Osborne, 2005; O'Connor, 2000). However, these strategies are likely better indicators and are seen as more ideal options in deciding on factor retention (Basto & Pereira, 2012; O'Connor, 2000; Velicer & Jackson, 1990).

Velicer’s MAP examines systematic and unsystematic variance with different numbers of factors (O’Connor, 2000; Velicer, 1976). Using principal components analysis, a partial correlation matrix is computed. Average values are squared, and factor retention will occur until the unsystematic variance is greater than the systematic variance (O’Connor, 2000). Parallel analysis also examines variance but compares the variance of the model to random variance (Horn, 1965; O’Connor, 2000). In this procedure, random data – matching in number of cases and variables to the study’s data – is used to compute eigenvalues and matrices. Factor retention will occur until the random data has a greater eigenvalue for the same number of factors compared to the target data set (O’Connor, 2000).

After examining the extraction criteria, researchers are recommended to extract factor models with one more and one less factor than the criteria suggest. Overall, researchers seek to find a solution with strong factor loadings (i.e., above .30), very few or no loadings of items on multiple factors, and factors with more than three items (Costello & Osborne, 2005).

EFA versus PCA. There is debate among researchers of whether principal components analysis (PCA) is a factor analytic procedure (e.g., Osborne & Banjanovic, 2016). As PCA is a commonly used data reduction technique, it should be mentioned. Moreover, PCA is a default among common statistical software programs (Costello & Osborne, 2005). While EFA and PCA both reduce data and extract either latent components or factors, there are some key distinctions. PCA assumes no error variance in contrast to EFA which accounts for and separates the unique and shared variances of items. As such, PCA is not likely as generalizable as it is specific to the sample (Osborne & Banjanovic, 2016). Although many researchers use PCA, this methodology is not appropriate for all data, especially with uncorrelated factors and moderate communalities (i.e., common variance of variables) as this could result in the components accounting for more variance (Costello & Osborne, 2005; Osborne & Banjanovic, 2016; Yong & Pearce, 2013).

Confirmatory Factor Analysis

Confirmatory factor analysis (CFA) is another statistical analysis for data reduction, with some key differences. In EFA, the latent variables are unknown, and researchers use methods (e.g., rotation, extraction) to explore a factor structure that best fits the sample. However, with CFA, there are no exploratory methods. The researcher pre-specifies aspects of the factor model (e.g., number or factors, pattern of factor loadings) and examines how well the model fits their data (Brown, 2015; Floyd & Widaman, 1995; Gerbing & Hamilton, 1996). In order to make these a priori specifications, there should be strong research and theoretical evidence to support the model. CFA often uses the same methods of estimation and indicators of a strong solution (e.g., ML, high factor loadings). However, since factor loadings are fixed – with cross loadings usually set to zero – there is no rotation and factor loadings produced by the CFA will often be higher than those within the EFA. CFA provides a standardized solution, however, there are many unstandardized aspects such as latent variables or estimates (e.g., standard errors, significance testing). In all, EFA and CFA are both used to test models but differ in how they should be used – typically with EFA occurring first as a means to explore a factor structure, and CFA occurring after there is a stronger theoretical and empirical basis (Brown, 2015). It is important to know the distinction between these two analytic methods, as some researchers may consider their analyses as CFA, when they may in fact be partially EFA (Gerbing & Hamilton, 1996).

Role of Factor Analysis in Scale Development and Validation

Both EFA and CFA have utility in assessing the psychometric properties of assessments. As EFA is typically conducted first, it is often used to identify latent factors from items on an assessment (e.g., rating scales). CFA is commonly used to verify factors and item loadings with different populations (Brown, 2015; Floyd & Widaman, 1995). For example, it would be important for a clinician or researcher to know if a specific assessment is valid and generalizable to use with varying ages, races, ethnicities, differential diagnoses, etc. (Floyd & Widaman, 1995). Using latent variables and factor loadings, researchers can make decisions about if and how strongly existing items accurately represent the measured construct. Further, factor analytic methods contribute to how an assessment is scored. Factors may be validated and form subscales, or a higher-order factor may support the use of one total score. It can also be used to support the reliability of the instrument and its scores, beyond Cronbach's alpha statistic (Brown, 2015). In all, factor analysis can contribute important psychometric information in scale development and validation. It is an important tool to determine whether the assessment measures the intended constructs and to examine the generalizability of the measure to samples beyond that of the normative group (Brown, 2015; Floyd & Widaman, 1995).

Specifically relevant to the current project, researchers use CFA to assess how well a factor structure may fit across different samples that vary in some systematic and potentially important way. Thus, providing psychometric evidence for use and similar interpretation of the instrument across particular populations. For example, March et al. (1999) used CFA to examine the fit of the hypothesized factor structure of the Multidimensional Anxiety Scale for Children (MASC) – originally standardized in a normative school-aged sample – in a sample of children with ADHD (i.e., no anxiety-based selection). Further, Benuto et al. (2020) used CFA to examine the fit of the Beck Anxiety Inventory (BAI) in a sample of Latinx patients. The BAI was developed with a sample of adults characterized as "psychiatric outpatients," without reported information regarding ethnicity (Beck et al., 1988). In both cases, authors intentionally selected a sample with specific characteristics (e.g., diagnosis, ethnicity) to examine how well the factor structure fit the intended population.

When examining ASD-specific instruments, researchers may choose to include additional cases with differing diagnoses to increase the sample's variability or range of performance. This tends to be done in research contexts where restricted range is suspected to be an issue that could lead to reduced factor complexity in more homogenous samples or to rule out such a possibility (see Gaskin et al., 2017). For example, researchers may assess if this alters the factor structure in a meaningful way or if the hypothesized structure still fits due to symptom overlap with other neurodevelopmental and psychiatric disorders (Bishop et al., 2016).

In a CFA context, researchers have used factor analytic procedures with samples intentionally inclusive of a wider variety of case types, characteristics, diagnoses, etc. For example, Frazier et al. (2008) examined the fit of the Autism Diagnostic Interview – Revised (ADI-R) in a sample that included individuals with confirmed ASD and cases of suspected ASD. This sample's diagnoses, based on the ADOS and ADI-R, respectively, included autism (82.8%; 73.5%), PDD-NOS (15.3%; 19.3%), and no diagnosis (1.9%; 7.2%). Results generally supported a two-factor model, consistent with the DSM-5 conceptualization of ASD (Frazier et al., 2008). Additionally, Uljarevic et al. (2020) examined the Research Domain Criteria model using select items from the Social Responsiveness Scale – Second Edition (SRS-2) in a broad sample that included participants who were either typically developing, diagnosed with ASD, siblings of those with ASD, or diagnosed with other neurodevelopmental or psychiatric disorders. The authors indicated that the inclusion of individuals with both “normative (33.8%) and atypical (66.2%) development” increased the variability of the sample, specifically regarding social constructs (Uljarevic et al., 2020; p. 1252). Uljarevic et al. (2020) reported that a three-factor structure had an adequate model fit while the four-factor structure had a superior fit. However, their approach was highly specialized and guided by an a priori theoretical model of social constructs. Thus, its findings are not readily comparable with prior studies.

In an EFA context, researchers have also employed tactics to purposely increase the variability of their samples. Researchers can select samples with variable diagnoses to reflect a population with ASD and other comorbid disorders to provide evidence of a measure’s validity within this population. For example, Bishop et al. (2016) chose a broader sample – inclusive of ASD (50%) and non-ASD (50%; e.g., ADHD, mood or anxiety disorders, ID, etc.) diagnoses – in their examination of the factor structure of the ADOS-2 Module 3. Authors highlight the overlap of symptoms or behavioral characteristics between ASD and these disorders, and that using such a mixed sample may elucidate distinctions within the constructs (Bishop et al., 2016). Further, Magyar et al. (2012) used both EFA and CFA to examine the validity of the SCQ in a sample with Down Syndrome (DS), with and without ASD. The SCQ is a widely used screening measure and individuals with DS appear to be at increased risk for an ASD diagnosis. Results supported a two-factor structure representative of ASD diagnostic constructs (i.e., social communication, stereotyped behavior and unusual interests), providing preliminary support for the use of the SCQ with individuals with DS (Magyar et al., 2012). Additionally, Kidd and colleagues examined the factor structure of the SCQ and SRS-2 in a sample with Fragile X Syndrome (FXS) – reported to be frequently co-morbid with ASD in addition to having overlapping features. In their sample, 44% of FXS cases also had ASD, while 56% of FXS cases did not. This sample also varied in terms of intellectual functioning and language levels (e.g., approximately 70% of the comorbid FXS with ASD group had moderate to severe/profound ID; Kidd et al., 2019). The level of cognitive functioning is important to note because it further contributes to the diversity and variability of the sample, beyond an ASD diagnosis. Specific to the GARS, Volker et al. (2016) explored the factor structure of the GARS-2 in a sample with both ASD cases (50%) and cases with other significant developmental disabilities (50%). This sample yielded a factor structure that was very similar to previously published work with an ASD-only sample (i.e., Lecavalier, 2005), with only a small number of low-loading items deviating. In this combined sample, researchers also reported low cognitive functioning – the mean IQ for the sample was 60.61 (*SD* = 19.61; Volker et al., 2016).

This selection of research demonstrates the use of factor analysis of instruments intended to measure ASD constructs using more diversified samples that include ASD and other diagnoses – those that may commonly co-occur or have overlapping features with ASD. This type of case selection can be used for a variety of purposes including model generalization in other populations, intentionally supplementing ASD cases with other case types to reflect the potential range and variability of scores more fully, to counter possible restrictive range resulting from diagnostic-specific homogeneity, etc. Increasing variability could reveal a more complex factor structure and/or result in a more stringent test of model fit than in a more homogeneous ASD-only sample. In study two, the ASD sample in the CFA will be supplemented with additional non-ASD developmental disability cases. This mixed sample will better reflect the screening conditions in which the measure would typically be used and to provide a more stringent test of model fit.

Clinical Discriminant Validity in Scale Development and Validation

Clinical discriminant validity involves evidence of whether, and to what extent, an assessment measure can discriminate between groups (e.g., diagnostic samples) known to differ on the construct of interest. Such evidence can consist of mean differences between groups on the measure of interest, categorical accuracy of group classification results based on the proposed cut score, or other similar approaches to mapping results of the assessment measure upon groups known to differ on the target construct. Psychometric concepts, such as the sensitivity and specificity of a measure, convey important information regarding how well a measure can accurately identify those who are at risk for or have a disorder and those who are not at risk or do not have the disorder, respectively (Kuriakose & Shalev, 2016). Other concepts important to understanding the meaning of sensitivity and specificity involve ultimately classifying a case as either a true positive (i.e., a person being accurately identified with a disorder by the test), true negative (i.e., a person accurately identified by the test as *not* having the disorder), false positive (i.e., a person inaccurately identified by the test as having the disorder) or false negative (i.e., a person with the disorder who was *not* identified by the test as having it). The number and percent of each type of case in a study sample are typically represented in a 2 x 2 table (reflecting predicted diagnostic status according to the test result versus actual known diagnostic status established independently of the test). Sensitivity is calculated as the number of true positives divided by the sum of true positives and false negatives. Specificity is calculated by dividing true negatives by the sum of true negatives and false positives (Bandalos, 2018).

In order for a classification to be made in practice, assessment measures often involve use of a cut score to a convey classification results (e.g., on the GARS-3, an individual with a score above 70 is deemed “very likely” to have ASD). In research, different cut scores might be examined to identify different rates of sensitivity and specificity for a sample. It is important to note that in raising the cut score, a researcher may risk increasing false negatives, while lowering the cut score may lead to increasing false positives (Bandalos, 2018). Importantly, the sensitivity and specificity of a particular cut score may depend on the characteristics of the groups or samples being compared. Thus, different cut scores may be warranted, depending on the characteristics of the samples involved. For example, when using the SCQ with young children, Wiggins et a al. (2007) suggested using a different cut score when differentiating ASD from other developmental disorders (i.e., ≥ 11) – as opposed to the cut point (i.e., cut score of ≥ 15) recommended by test authors based on results from a ROC curve analyses comparing a PDD sample to a non-PDD sample with other psychiatric diagnoses (Berument et al., 1999).

Typically, when using a level 2 screener, there is already some data that pointed a clinician to further evaluation. With ASD, a clinician might use the GARS-3 to follow up with any indicators or concerns. While a level 1 screener like the M-CHAT might have high sensitivity to highlight any concerns or atypical development, a level 2 ASD screener should have high specificity in order to correctly identify those who do *not* have ASD (Kuriakose & Shalev, 2016; Lalkhen & McCluskey, 2008). Clinicians need accurate assessment measures when they gather data. A measure with a high rate of false negatives could lead to a child not receiving appropriate services based on inaccurate results, thus highlighting the importance of classification accuracy (Bandalos, 2018).

Testing and Assessment Standards

In best practice standards for educational and psychological testing, the psychometrics (e.g., validity, reliability) should be strong and support the use of the instrument for the intended population. Specific to the focus of this research, the validity, or how much support a test has for its intended use, is considered “fundamental” when creating and using assessments (American Educational Research Association [AERA], APA, & National Council on Measurement in Education [NCME], 2014, p. 11). Validity of a test can be assessed by looking at its content (e.g., Do the items represent the intended construct?), response processes (e.g., How are raters responding to the instrument?), internal structure (e.g., What are the relationships between the items and/or components? How do they relate to or represent the intended construct?), in addition to convergent or divergent validity, criterion-related validity, and generalizability. Specific to internal structure validity, best practice standards suggest that there should be evidence of the internal structure if the instrument/its score is comprised from interrelated items (Standard 1.13; AERA, APA, & NCME, 2014). Additionally, there should be evidence to support subtests and interpretations, specifically by providing evidence of having multiple, reliable scores (Standard 1.14; AERA, APA, & NCME, 2014). Further, evidence should be provided for use of interpretation at the item-level or of groups of items, especially if suggested by the test author (Standard 1.15; AERA, APA, & NCME, 2014). A factor analysis is one type of analysis that would provide evidence for the internal structure of the assessment and whether the test is comprised of one or multiple subtests/factors (AERA, APA, & NCME, 2014).

When strong validity is not presented, there may be misuse of and inaccurate interpretations from the instrument (AERA, APA, & NCME, 2014). Therefore, examining the internal structure through factor analysis is pertinent to the use and interpretation of an assessment, like the GARS-3. The GARS-3 manual reports some details of an exploratory factor analysis, but more detailed information and further independent analysis with another sample would better support these best practice standards. Additionally, evidence to support its validity in discriminating between different developmental disorders is key to score interpretation and accurate clinical utility.

Factor Analyses of the GARS

GARS and GARS-2

To date, there are four studies that examine the factor structure of the first two versions of the GARS. Lecavalier (2005) examined the internal structure of the original GARS through exploratory factor analysis. The sample consisted of 284 students (mean age = 9.3), who were majority White and male (i.e., 91.7% White, 79.2% male) and whose parents and/or teachers completed GARS ratings. Lecavalier used PCA with direct oblimin rotations to examine the structure of the GARS compared to the three published behavioral subscales (i.e., Stereotyped Behaviors, Communication, and Social Interactions). Results indicated communalities for the 42 items ranged from .17 to .58, the Kaiser-Meyer Olkin test was .88 (i.e., high or “meritorious”; Kaiser & Rice, 1974), and the Bartlett test was significant – indicating that PCA was an appropriate method for the data (Lecavalier, 2005). The scree plot yielded three clear factors that aligned with the general DSM-IV criteria, although only explained 37.6% of the variance – a low value. The first factor, Repetitive Behavior, explained 23.8% of the variance and from the original GARS subscales, predominantly included items from the Stereotyped Behavior subscale, but also included 2 items from Communication, and 5 items from Social Interactions. Factor II, named Social Interaction, explained 7.2% of the variance and included 9 of 14 items from the original Social Interaction subscale, 3 items from Communication, and 1 item from Stereotyped Behavior. The third factor, Communication, consisted of items that were all from the original Communication subscale. Overall, Lecavalier’s results concluded that only 74% of the items in the study loaded on the same factor as the published GARS model (Gilliam, 1995), and that almost half (i.e., 48%) loaded onto the Repetitive Behavior factor (Lecavalier, 2005). Lecavalier also reported that most of the items had “acceptable” loadings on only one factor and reported high correlations between the three published behavioral subscales (ranging from .48 - .62; Lecavalier, 2005).

The Lecavalier article was the sole published independent factor analysis of the original GARS, and subsequent researchers have focused on the second edition of the GARS. Pandolfi et al. (2010) examined the internal factor structure of the GARS-2 using exploratory and confirmatory factor analytic methods. Their sample included a total of 1,129 individuals aged three through 22 years, the majority of whom were a part of Gilliam’s standardization sample. An additional 22 individuals were included in the sample after the publication of the GARS-2. Raters included parents, close relatives, teachers, and other professionals. Researchers examined the polychoric correlations from the data finding that the assumption of normality was “tenable for each item pair” in addition to no observed pattern of empty cells (Pandolfi et al., 2010, p. 1121). In their exploratory factor analysis, 3-, 4-, 5-, and 6-factor solutions were extracted using varimax and promax rotations. The authors denoted that promax was the preferred rotation as the measured ASD subscales or factor-based constructs were likely to be correlated. In their analysis, coefficients greater than or equal to .32 were meaningful as this indicated a level of item variance due to the underlying construct/factor. Results showed a 4-factor solution that explained 38.63% of the variance was most interpretable: Factor I was named Stereotyped/Repetitive Behavior and explained 11.33% of the variance and contained 11 of 14 items of the corresponding GARS-2 subscale; Factor II was titled Stereotyped/Idiosyncratic Language which explained 8.25% of the variance; Factor III was named Word Use Problem and explained 5.49% of the variance; and Factor IV was labeled Social Impairment and explained

13.53% of variance and contained 9 items from the Social Interaction GARS-2 subscale (Pandolfi et al., 2010).

Using confirmatory factor analysis, Pandolfi et al. (2010) also examined two models, using Diagonally Weighted Least Squares estimator – the published model from the original GARS-2 in addition to the four-factor model found with their exploratory analysis. Authors used fit indices to assess the fit of the model which included the Satorra-Bentler chi-square statistic, the Root Mean Square Error of Approximation (RMSEA) and the Comparative Fit Index (CFI). Findings from the CFA supported the study's four-factor solution, however, did not support Gilliam's theoretically derived model.

Pandolfi et al. (2010) discussed some of the potential problems with the GARS-2. The authors criticized the relevancy of some items to the core constructs of ASD – the items that did not load onto their theoretical construct might be correlated, but not definitive of the construct. For example, establishing eye contact might correlate with restrictive and repetitive behaviors, but may be more indicative of social impairment. Authors also criticized the use of "double-barreled" items or items that ask raters to attend to different behaviors. This makes it difficult for raters to be consistent in their responses, affecting accuracy of the instrument, as raters may be responding to different aspects of the same item (Pandolfi et al., 2010).

In a study by Volker et al. (2016), researchers examined the factor structure of the GARS-2 using teacher ratings. Their sample included 240 individuals whose mean age was 9.5, mean IQ was 60.61, and included those with ASD and other developmental disabilities (e.g., ID). The majority of the sample was Caucasian (79.58%) and male (78.75%). Teachers and staff members, familiar with ASD characteristics, rated the sample of 240 individuals. Researchers attempted to have individuals rated by the staff member who knew them best, but also prioritized independence in their ratings (i.e., various types of staff members in each classroom rated each individual and completed the scales without access to the individual's record). In the EFA, researchers used the polychoric correlation matrix with PAF and a promax rotation, and the maximum correlation method to estimate item communalities. The results suggested a three-factor structure which the researchers named and conceptualized as Stereotyped and Repetitive Behaviors, Social Avoidance and Withdrawal, and Atypical Language and Communication. Factor I (Stereotyped Behaviors) contained 18 items, including 12 of the original 14 items of the Stereotyped and Repetitive Behaviors subscale, 2 from the original Communication subscale, and 4 from the original Social Interaction subscale. Authors noted that the two ladder sets of items either directly or indirectly related to stereotyped or repetitive behaviors. Factor II (Social Avoidance and Withdrawal) consisted of 16 items and included 8 of the original 14 items from the Social Interaction subscale, 6 from the original Communication subscale and 2 from the original Stereotyped Behaviors subscale. Authors concluded that the ladder items loaded highly on this factor due to their relatedness to social avoidance or withdrawal. Factor III (Atypical Language and Communication) consisted of 8 items – 6 items from the original Communication subscale in addition to 2 from the original Social Interaction subscale. Researchers criticized the specificity of the measure, suggesting that the GARS-2 may be under-classifying cases. Although results suggest retention of the same number of factors as the published model, there were discrepancies in convergent item loadings such that 38% of items did not load where predicted. Researchers also looked at the clinical utility and discriminant validity of the measure between groups with ASD and non-ASD developmental disabilities. Their results indicated that 34.71% of ASD cases did not meet the cutoff score and were classified as non-ASD, although completed by special education teaching staff who are generally more knowledgeable and experienced with ASD compared to the typical population (Volker et al., 2016).

Volker et al. (2022) revisited the GARS-2 factor structure using raters from a large special education agency, who were primary or support teaching staff. The sample included 216 individuals between the ages of three and 21 years, who had a diagnosis of autistic disorder or PDD-NOS based on the DSM-IV-TR or received special education services through a classification of autism. Researchers analyzed data using EFA with PAF and the polychoric correlation matrix. Researchers retained a three-factor solution, using a promax rotation, which accounted for 73.54% of the variance. Factor I (Stereotyped and Repetitive Behaviors) contained 19 items and included 12 of the 14 items on the original GARS-2 Stereotyped Behavior scale. Factor II (Social Avoidance and Withdrawal) contained 17 items and included 8 of 14 of the original items on the GARS-2 Social Interaction subscale. Factor III (Atypical Language and Communication) contained only 6 items, all from the original GARS-2 Communication scale. In addition to these findings, researchers also examined results in the context of previous factor analyses. Specifically, Volker et al. (2022) looked at item agreement between studies and found 92.86% agreement with Volker et al. (2016), 85.71% agreement with Lecavalier (2005), and 78.57% agreement with Pandolfi et al. (2010). Taken together, independent research is important for assessing the generality of findings reported in test manuals across different samples, populations, raters, contexts, etc.

GARS-3

Since previous research studies involving the GARS and GARS-2 have yielded factor structure and item loading discrepancies between the publishing author and outside researchers, it is important to look critically at existing data reported in the manual. The factor analysis reported in the manual is the only factor analytic data of the GARS-3. The manual reported using the sample of 1,859 individuals diagnosed with autism. Researchers conducted an EFA using ML and a promax rotation. Results yielded a six-factor solution, as represented by the current six subscales of the GARS-3 (Gilliam, 2013).

The test author reported information regarding standards used to assess the appropriateness of the data for EFA. The Bartlett Test of Sphericity was reported as significant ($\chi^2 = 66272.7$, $p < .001$), indicating suitability for EFA (Gilliam, 2013). The Bartlett test indicates that the correlation matrix is not an identity matrix, which is not a likely occurrence (Bartlett, 1950; Raykov & Marcoulides, 2008). Further confirming that the correlation matrix was not an identity matrix, the author reported that most correlations were at or above .3 (Gilliam, 2013). The Kaiser-Meyer Olkin (KMO) Measure of Sampling Adequacy was used and was reported as .95. The manual indicated this was a high value as the recommended value is .6 (Gilliam, 2013). The KMO can range from zero to one with higher values indicating better sampling adequacy or more shared variance among the variables in the matrix (Kaiser, 1970; Kaiser & Rice, 1974). There is a series of KMO standards that indicate recommended values which Kaiser & Rice (1974) labeled with qualitative markers. A KMO of .6 would be labeled as "mediocre," while higher values such as .8 or .9 would be considered "meritorious" and "marvelous," respectively (Kaiser & Rice, 1974).

As previously indicated, the test author used ML and a promax, or oblique, rotation in the GARS-3 EFA (Gilliam, 2013). An advantage of using ML is that it is an indication of fit (Osborne & Banjanovic, 2016). It is ideally used when data are normally distributed, however, the author does not report if the data met this normality assumption (Gilliam, 2013; Osborne & Banjanovic, 2016). If the normality assumption is not met, PAF is typically used (Costello &

Osborne, 2005; Yong & Pearce, 2013). A promax rotation is typically preferred when the factors are hypothesized to be related – as in this case with constructs of ASD (Costello & Osborne, 2005; Pandolfi, 2010).

The reported factor analysis yielded a six-factor solution. The manual reported the percentage of variance accounted by each factor as follows: the Social Interaction factor described 46% of the variance, the Restrictive/Repetitive Behaviors factor described 18% of the variance, the Social Communication factor described 11% of the variance, the Emotional Responses factor described 9% of the variance, the Cognitive Style factor described 8% of the variance, and the Maladaptive Speech factor described 4% of variance. The manual reported that the scree plot was examined in determining the number of factors to retain (Gilliam, 2013). The scree plot is a graph of the factors' eigenvalues and researchers may reference the "break" or "point of inflexion" in the graph to gauge the number of potential factors (Costello & Osborne, 2005; Yong & Pearce, 2013). However, using a scree plot is considered more of a "rule of thumb" analysis and research suggests this method lacks accuracy (O'Connor, 2000). Eigenvalues were reported only for the six retained, rotated factors (Gilliam, 2013). While all eigenvalues were greater than one, there was no indication whether the author also considered the Kaiser criterion in factor retention (i.e., retention of factors whose eigenvalues are greater than one; Costello & Osborne, 2005; Gilliam, 2013). Although researchers do not want to solely retain factors based this criterion, as it is reported as inaccurate and unreliable, knowledge of factor eigenvalues would provide another indicator in the factor retention decision (Costello & Osborne, 2005; O'Connor, 2000). Further, researchers did not report using other methods of factor selection, such as using parallel analysis and Velicer's MAP – often recommended as the ideal strategies (Basto & Pereira, 2012; Costello & Osborne, 2005).

Overall, the information reported in the GARS-3 manual regarding the EFA was limited. There are additional information and indicators needed to inform the validity of the analysis. As mentioned, more ideal strategies such as parallel analysis and Velicer's MAP would provide greater accuracy in the decision of factor retention (Basto & Pereira, 2012). Further, information about the normality of the data is needed in deciding the appropriate estimation method (e.g., ML or PAF; Osborne & Banjanovic, 2016). Additionally, it would be beneficial to know of any other factor solutions considered. Within EFA, researchers use multiple indicators such as the scree plot, Kaiser criterion, parallel analysis, and Velicer's MAP to inform solutions that will be interpreted (e.g., looking at solutions with a number of factors both greater and less than the number of factors based on the indicators; Costello & Osborne, 2005; O'Connor, 2000). From these solutions, researchers can decide upon the most interpretable solution that best fits the data. Ideally the factors have strong loadings greater than .3, few or no cross loadings of items, and strong factors with more than three items (Costello & Osborne, 2005).

Although the GARS-3 did not report the type of correlation matrix used, the GARS-3 item scaling is best interpreted as ordinal data (Gilliam, 2013). The 4-point rating scale should be considered ordinal as they consist of four ordered categories that describe the behavior. There are relationships between the scores, but the scores may not be equidistant from each other – in the GARS-3 rating scale, there is likely not a standardized and/or equal difference between categories of "not at all," "not much," "somewhat like," and "very much like" when rating the behavior of an individual (Salvia et al., 2017). Covariations with ordinal data are most appropriately assessed using the polychoric correlation coefficient, thus, the scale might be best interpreted as four ordinal categories within a polychoric correlation matrix, rather than continuous categories associated with Pearson's correlation matrix. Given its capability of estimating the more continuous nature of the underlying construct, the polychoric correlation matrix would likely yield correlations greater than or equal to the Pearson correlation matrix (Basto & Pereira, 2012). While the type of correlation matrix was not indicated, Pearson's *R* correlations are commonly used – as they might be the only option for many statistical software programs – and if researchers do not mention a specific correlation, it is likely that Pearson's *R* was used (Basto & Pereira, 2012; Gilliam, 2013).

Need for the Present Study

This research provides needed independent research regarding the validity of the GARS-3 for use in research and practice. As mentioned, the GARS-3 is used in research to verify sample diagnosis, measure severity of symptoms, assess ASD-related symptoms, to validate other ASD measures, and to measure treatment change (e.g., Alsaedi et al., 2020; Duffy et al., 2017; Eskow et al., 2019). Previous and current versions of the GARS are also cited as being widely used in schools (e.g., Aiello et al., 2017 found the GARS-2 was the most frequently used ASD rating scale; Benson et al., 2019 found that the GARS-3 was the third most used ASD-related assessment tool). Researchers and clinicians need to be confident about the measures they use, therefore, more support for the psychometrics of the GARS-3 is needed.

As noted, the only existing factor analytic data for the GARS-3 is reported in the manual with the standardization sample (Gilliam, 2013). There are several reasons why additional factor analyses are essential to understand the number and types of factors with the GARS-3. First, individual samples involve sampling error which can influence the pattern of results in EFA (Brown, 2015). By replicating results in a different sample, researchers can account for sample variation and have increased confidence in a factor structure. If the factor structure is not generalizable, there may be evidence to support important moderator variables that influence the outcome of results. Second, samples may vary systematically due to selection criteria or methods. In the standardization sample of the GARS-3, raters were recruited online, some ratings were completed online, and limited diagnostic information may have been involved (Gilliam, 2013). Because of potential meaningful differences of these aspects, other samples may differ in any or all these ways. Third, differences in raters have the potential to yield different patterns in ratings and/or differences in relationships between items. In the GARS-3 standardization sample, variation across raters may have been averaged due to the wide variety of rater types and settings (e.g., parents, teachers, psychologists, educational diagnosticians, speech clinicians, teaching assistants, other school personnel). Fourth, previous research studies involving the GARS and GARS-2 have yielded factor structure discrepancies between the publishing author and outside researchers (e.g., Pandolfi et al., 2010; Volker et al. 2016; Volker et al., 2022). Significant differences were found in both the number of factors retained and regarding which items should be assigned to specific factors/subscales. Thus, it is important to look critically at the generalizability of the GARS-3 factor structure as the literature suggests evidence to not solely consider the published factor structure. Finally, the EFA reported in the GARS-3 manual (Gilliam, 2013) had limited details, used a factor analytic procedure that is not robust to the violation of normality assumption (i.e., maximum likelihood), and did not involve what would be considered best available methods for EFA (e.g., reliance on the scree plot and Kaiser criterion and did not include parallel analyses or Velicer’s MAP; Basto & Pereira, 2012; O’Connor, 2000).

Additionally, there is also a need to examine the ability of the GARS-3 to discriminate between samples. The manual reported high levels of sensitivity (Autism Index 6 = .96) and specificity (Autism Index 6 = .84) when discriminating between individuals with ASD and individuals with non-ASD diagnoses (i.e., ID, deaf, blind, ADHD, ED/BD, LD, and physical/health impairment). However, when looking specifically at individuals with diagnoses that may share similar features with ASD (e.g., ID), mean differences between Autism Index scores were significant, but scores of individuals with ID would still be deemed as having a high likelihood of ASD (cut score of 70; Autism Index 6 $\bar{X} = 87$; Gilliam, 2013). Additionally, independent research with previous versions of the GARS has highlighted high rates of false negatives. For example, with the GARS, South et al. (2002) reported that 52% of the ASD sample was inaccurately classified and with the GARS-2, Volker et al., (2016) indicated 34% of the ASD sample was not accurately categorized. Classification accuracy in screening measures – especially in level two screening – is critical as it provides evidence to support a diagnosis *or* evidence that the individual needs further ASD-specific evaluation. Since individuals with ASD and other developmental disabilities that require significant support might have similar features (e.g., repetitive behaviors, communication difficulties; APA, 2013; Matson & Shoemaker, 2009) – and may often be rated by this measure – it is important to provide evidence of the effectiveness in clinical discriminant validity.

Given the necessity to address psychometrics of the GARS-3 with independent research, the present study conducted an exploratory factor analysis, a confirmatory factor analysis based on findings, and analyses to determine clinical discriminant validity using ASD and non-ASD DD samples rated by special education staff. This study will contribute to the literature by providing needed independent evidence of validity in a sample of raters and individuals differing from the normative sample, examining the generalizability of the GARS-3 factor structure. This sample is unique as it is comprised of special education staff members such as teachers, teacher aides, physical therapists, and occupational therapists. In contrast, the GARS-3 normative sample was a combination of several types of raters (e.g., parents, teachers, psychologists; Gilliam, 2013). Additionally, individuals within this sample have a high prevalence of functional impairment and receive educational and behavioral services/support from a center-based special education agency. Given the rated individuals and the group of raters, these specific demographic changes may influence the relationships between subscale items and also, provide an opportunity to assess clinical discriminant validity in different samples of developmental disabilities. In all, this study provides essential independent psychometric information for the GARS-3.

The following research questions were investigated within this dissertation, pertaining to exploratory factor analysis (i.e., research questions one through four), confirmatory factor analysis (i.e., research questions five through seven), and clinical discriminant validity (i.e., research question eight).

Study One Research Questions

Research Question 1

When students with ASD are rated by special education teaching staff with the GARS-3, how many potentially interpretable factors are present and should be considered for retention?

Research Question 2

When students with ASD are rated by special education teaching staff with the GARS-3, how many factors should be retained to yield the most interpretable factor solution for the GARS-3?

Research Question 3

When students with ASD are rated by special education teaching staff with the GARS-3, are there substantive correlations between at least some GARS-3 factors within the most interpretable factor structure?

Research Question 4

When students with ASD are rated by special education teaching staff with the GARS-3, how does the six-factor EFA solution correspond to the six GARS-3 subscales proposed by the author (Gilliam, 2013)?

Study Two Research Questions

Research Question 5

When students with ASD and non-ASD developmental disorders are rated by special education teaching staff with the GARS-3, does the interpretive model proposed by the GARS-3 test author produce a reasonable fit to the confirmatory sample covariance matrix?

Research Question 6

When students with ASD and non-ASD developmental disorders are rated by special education teaching staff with the GARS-3, does the retained factor solution from the study one EFA produce a reasonable fit to the confirmatory sample inter-item covariance matrix?

Research Question 7

When students with ASD and non-ASD developmental disorders are rated by special education teaching staff with the GARS-3, and the GARS-3 author-proposed model and the EFA-generated model from study one are compared, does one model show evidence of better fit to the confirmatory sample inter-item covariance matrix?

Study Three Research Question

Research Question 8

When students with developmental disabilities, in a center-based special education setting, are rated by special education teaching staff using the GARS-3, how well does the GARS-3 discriminate individuals with ASD from individuals with other developmental disabilities that require substantial support?

CHAPTER THREE: METHOD

This dissertation consists of three inter-related studies. All three studies addressed aspects of the broader construct validity of the GARS-3 (i.e., study one and two related to internal structure validity [consistency of relationships internal to the test with the intended constructs or model], study three examined clinical discriminant validity [extent to which scores derived from the instrument differentiate between know groups in a manner consistent with the intended construct]). The first study consisted of an exploratory factor analysis (EFA), which addressed research questions one through four. The second study consisted of a confirmatory factor analysis (CFA) and addressed research questions five through seven. Finally, study three evaluated the clinical discriminant validity of GARS-3 scores by testing mean score differences between clinical groups (ASD versus other developmental disabilities [DDs] that require substantial support), examining screening utility of the cut scores proposed for the GARS-3 Autism Index composite, and exploring the classification accuracy of all possible cut scores to determining an optimum cut score for differentiating between the groups involved in the study. This study addressed research question eight.

Research Design

The three interrelated studies focused on important aspects of instrument validation (i.e., internal structure validity, model fit, and clinical discriminant validity of the GARS-3) in the context of ASD and non-ASD DD samples rated by special education staff. In terms of design elements, these studies are all cross-sectional, in that they assessed data at one point in time; correlational, in that they assessed associations between variables at that time point; and observational, in that they involved only measured variables and no manipulated (i.e., independent) variables (Kazdin, 2017). Study one (EFA) involved a single large sample of cases with ASD and study two (CFA) involved a second large sample of cases with ASD and non-ASD developmental disorders. Study three (clinical discriminant analysis) involved the comparison of two different, previously established, clinical samples (i.e., one with ASD and the second with other significant developmental disabilities). Studies one and two pertained to internal structure validity, and specifically, the factor structure of the GARS-3. Factor analytic techniques are multivariate in nature. They reduce observed inter-item correlations to one or more latent variables (i.e., factors). Such latent variables are intended to explain covariation among a set of observed variables (in this case, rating scale items). The name assigned to each latent variable is intended to reflect the underlying construct represented (i.e., the meaning of the factor). In scale development, an exploratory factor analysis can be used to initially examine underlying constructs and, further, to inform the scoring of the instrument (Brown, 2015; Floyd & Widaman, 1995). It is critical to replicate the factor structure across different independent samples to account for the possible influence of sampling error on the factor structure and assess the generalizability of the factor model (Brown, 2015; Floyd & Widaman, 1995). Through EFA, this study examined the factor structure of the GARS-3 in a sample of students with ASD as rated by special education staff members who know the students well. As part of an extension of the EFA, the consistency of factor-derived subscales with the six GARS-3 subscales proposed by the test author (Gilliam, 2013) was also examined. In study two, a CFA was conducted using a second sample from the same special education agency as in study one, including ASD and non-ASD developmental disorders. The CFA involved testing the fit of both the factor structure derived from the study one EFA and Gilliam’s proposed six-factor structure (Gilliam, 2013), and comparing the relative fit of the two factor models.

In addition to internal structure validity, this dissertation examined the clinical discriminant validity of the GARS-3 in study three. In this study, cases with ASD were compared to cases with other, non-ASD, developmental disabilities involving substantial impairment. They were compared in terms of between-group mean differences, classification accuracy based on the author-proposed cut score, and exploration of all possible screening cut scores using receiver operator characteristic (ROC) curve analyses (Metz, 1978) to determine the optimal cut score for the study sample. Classification accuracy is critical in establishing the practical utility of an instrument for identifying cases with and without a particular disorder (Bandalos, 2018). It involves evaluating the classification accuracy of cut scores, in particular samples, in terms of sensitivity (true positives / [true positives + false negatives]), specificity (true negatives / [true negatives + false positives]), positive predictive power (true positives / [true positives + false positives]), and negative predictive power (true negatives / [true negatives + false negatives]; Lalkhen & McCluskey, 2008). These results were compared to those reported in the GARS-3 manual for different groups and raters.

Data Collection (Extant Data Set)

Data for the three studies are part of a large existing program evaluation dataset from a special education agency in Western New York State. This agency specializes in the treatment of students with moderate to severe developmental disabilities. Cases involved in the present study required center-based services, indicative of substantial impairment and need for intensive supports that are not possible in a typical educational setting. Based on prior samples from this organization, general cognitive ability for the majority of students was in the intellectual disability range (e.g., Birnbaum, 2020).

These data were collected as part of annual program evaluations over a five-year period (2015-2020). Cases that occurred in two or more years (e.g., an individual was rated first in 2015 and then rated again by a different staff member in 2016), were randomly distributed between the samples for studies one and two. In such cases, any additional rating occurred one-to-two years after and was completed by a different staff member. All cases within a given factor analysis were unique and independent, however, some overlap in cases occurred between study one and study two. In order to create a sample size with enough power with factor analyses, and with a goal of *n* = 200 for the CFA, 18 unique cases with ASD diagnoses were selected for the study two sample, while the remainder were assigned to the EFA sample (i.e., *n* = 204). Thus, although some cases overlapped across the studies, the particular participant, rater, and time point combination was unique to each factor analytic study.

Data consisted of demographic information for each case (i.e., age, gender, ethnicity, diagnosis [ASD vs. non-ASD DDs], cognitive deviation quotient [most recent available estimate of general cognitive functioning], and a code for verbal status [presence of speech/spoken language versus no spoken language]), the 58 items of the GARS-3, the six subscale scores for the GARS-3, and the Autism Index 6 and Autism Index 4 composites for the GARS-3 (please refer to Appendix A for more information about the subscale and composite scores). Though several items make a reference to the use of signs, the instructions given on the rating form after the fourth subscale instruct the rater to indicate if the individual is “mute.” If answered yes, then the rater should not complete the two subsequent subscales (i.e., Maladaptive Speech, Cognitive Style). It was not possible to collect a case-specific indicator of socio-economic status. Based on prior samples from this organization, the agency reported 29-36% of students qualified for free and reduced lunch (Birnbaum, 2020). Privacy restrictions would not allow for reporting them in connection with individual cases.

Raters

Raters consisted of special education staff including teachers, teaching assistants. teacher aides, speech pathologists, physical therapists, and occupational therapists who work with the rated individuals. Selected measures are annually rated for each individual attending this special education agency. These raters are different from parents or mainstream teaching staff in a typical educational environment, because as employees working in an intensive special education setting, they tend to have more training and expertise in disorders such as ASD. However, because of the intensive behavioral orientation of the intervention program and the diversity of developmental disabilities represented, in most cases, raters were not aware of which students had a formal ASD diagnosis and which did not.

Procedures

The program evaluation was completed annually and involved each staff member completing a packet of ratings scales for an assigned student. For every year of the program evaluation, each student was rated by a different staff member. Because of the large special education team working with each center-based classroom, it was possible to achieve a near 1:1 correspondence between student and rater. Rating assignments were allocated to staff by the director of program evaluation, in consultation with lead teachers. At the time of data collection, each staff member completed five rating scales (in random, counter-balanced order) for an assigned student. Staff members were instructed to complete the rating scales in the order presented. In general, staff members who knew each student best were preferentially assigned for ratings, when possible. Each rater had known or worked with the student being rated for at least three months, and upwards of 28 months.

When rating packets were turned in by staff, they were immediately checked over for missing item ratings or unclear ratings (e.g., two numbers circled on the rating scale for the same item). Such omissions or errors were resolved by consulting the staff rater involved. Once ratings were determined to be complete, each rating scale was independently scored by two different, trained, program evaluation staff. Any score disagreements between the two scorers were resolved by an additional independent scoring of the instrument by the director of program evaluation. Once scored, all item ratings, subscale scores, and composite scores were double entered into a database by independent program evaluation staff members. Any data entry discrepancies were resolved by a third independent staff member who finalized the data set.

Measure

The measure used in each of the three studies, and the subject of validation across the three studies, is the Gilliam Autism Rating Scale – Third Edition (GARS-3; Gilliam, 2013). The GARS-3 can be used by parents, caregivers, and teachers as a level two ASD screening measure for individuals ages 3 through 22 years. Items are rated using a 4-point rating scale (i.e., 0 = "Not at all like the individual", 1 = "Not much like the individual", 2 = "Somewhat like the individual", and 3 = "Very much like the individual") in terms of how well they describe the person being rated. The instrument consists of 58 items divided among six subscales: Restricted/Repetitive Behaviors (13 items), Social Interaction (14 items), Social Communication (9 items), Emotional Responses (8 items), Cognitive Style (7 items), and Maladaptive Speech (7 items). A composite score, the Autism Index, can be calculated to indicate the probability of ASD. This Autism Index is a linear transformation of the sum of scaled scores from the corresponding subscales. If an individual is “mute” or not communicative, a rater completes four subscales to allow for calculation of the Autism Index 4. If the individual is verbal, a rater completes all six subscales to calculate the Autism Index 6. More in-depth information regarding the GARS-3 and its reliability and validity can be found in the Literature Review (see Chapter 2).

Inclusion Criteria

To be included in study one, a participant was (a) between the ages of three and 22 years; (b) have a clinical diagnosis of ASD under the DSM-5 (APA, 2013), previously diagnosed under DSM-IV-TR with autistic disorder or PDD-NOS (APA, 2000), or receiving special education services under the eligibility of autism spectrum disorder; (c) with sufficient verbal communication skills for completion of the GARS-3 items that assume an individual can speak (i.e., the GARS-3 record form asks the rater whether or not the rated individual is “mute” to determine if additional items should be completed), and (d) students who require center-based special education services.

Participants in studies two and three had much of the same inclusion criteria, specifically criteria a, c, and d from above. The clinical diagnoses in study two included criterion b above, but also included (e) individuals with other, non-ASD, developmental disorders; they require significant support and had a clinical diagnosis of another developmental disability as listed in the DSM-IV-TR (APA, 2000) or DSM-5 (APA, 2013), or receive special education services under a related eligibility category (e.g., Intellectual Disability). This group is characterized as having multiple moderate to severe delays across important domains of functioning (e.g., cognitive ability, adaptive behavior, language, motor, etc.). Because study three involved two samples, one sample met the complete criteria b, described above. The second group met criteria e from study two.

Study One: EFA

Research Questions, Rationale, And Hypotheses

Research Question 1. When students with ASD are rated by special education teaching staff with the GARS-3, how many potentially interpretable factors are present and should be considered for retention?

Research Question 2. When students with ASD are rated by special education teaching staff with the GARS-3, how many factors should be retained to yield the most interpretable factor solution for the GARS-3?

Rationale. Only one prior factor analysis of the GARS-3 has been reported. This prior analysis was exploratory and conducted by the test author on the standardization sample (Gilliam, 2013). There are several reasons why further factor analyses, especially EFAs, should be conducted. Characteristics of the sample (i.e., individual samples involve sampling error, which can significantly influence the results of exploratory analyses; methodology in selection of the sample) and its raters (e.g., possible greater variation in ratings with a wider variety of rater types including parents, teachers, psychologists, etc.) may vary across samples and therefore, should be replicated to increase confidence in their generalizability. Further, previous factor analyses conducted with the GARS and GARS-2 by independent researchers found significant divergencies in the resulting factor structure relative to the author-designated subscales—especially regarding which items should be assigned to which subscales (e.g., Lecavalier, 2005; Pandolfi et al., 2010). Thus, there is a need to verify results reported in the manual. Finally, the EFA reported in the GARS-3 manual was limited in detail (e.g., no inter-factor correlations were reported), used a factoring procedure that is not robust to the violation of normality assumptions [i.e., maximum likelihood] which are not likely to be met by ratings of an ASD sample, and did not involve what would be considered the best available factor selection methods for EFA (e.g., analyses reported in the manual relied on the scree plot and Kaiser criterion, but did not involve use of parallel analysis or Velicer's MAP; Gilliam, 2013).

The present EFA involved a relatively well characterized sample (i.e., verifiable diagnoses according to DSM criteria or special education classification) with raters from a special education agency that specializes in interventions for developmental disabilities and who know the students well. Further, ratings were performed using physical response booklets, data were collected in person, and strong data quality control measures were implemented (e.g., record forms immediately checked for missing responses and promptly corrected, rating forms independently scored by multiple trained scoring personnel, and data independently double-entered into the database). EFA methods for this study involved best practice methods (e.g., Basto & Pereira, 2012; Osborne & Banjanovic, 2016). This included procedures for factor analyzing ordinal data (i.e., inter-item polychoric correlation matrix instead of Pearson's correlation matrix), use of a factoring procedure that is robust to violations of normality assumptions (i.e., principal axis factoring [PAF] in contrast to maximum likelihood), use of parallel analysis and Velicer's MAP as part of factor selection decisions, use of an oblique rotation to account for factor correlations, and use of multiple independent researchers familiar with ASD to independently interpret the various candidate factor solutions. Thus, confidence in the results, in terms of number of factors and their interpretation, are strong for the present sample, as shortcomings of the EFA reported in the GARS-3 manual were addressed.

Research Question 3. When students with ASD are rated by special education teaching staff with the GARS-3, are there substantive correlations between at least some GARS-3 factors within the most interpretable factor structure?

Hypothesis 3. Within the most interpretable factor structure, correlations between at least some factors will be ≥ .30.

Rationale. It was anticipated that the factors will be substantively correlated. Although there is no clear standard for an inter-factor correlation, there is consensus that when the inter-factor correlations are at or near zero, an orthogonal rotation is most appropriate. However, most authors do not give clearly defined cut-off value for when inter-factor correlations are substantive. In general, if inter-factor correlations fall below .30, consideration should be given for an orthogonal rotation if it simplifies the analysis without significant distortion (Nunnally & Bernstein (1994) as cited in Pett et al. [2003]). Otherwise, the obliquely rotated solution should be retained. Theoretically, it would be expected that different components of a diagnosis would be correlated. Similarly, Norris et al. (2012) examined different factor models of ASD symptomatology and found varying inter-factor correlations for the DSM-IV (three-factor) and DSM-5 (two-factor) models, with correlations ranging from .75 to .92 using ADOS Module 1 data and from .33 to .93 using ADOS Module 3 data. Other major ASD rating scales, such as the SRS (e.g., Frazier et al. 2012; Nelson et al., 2016) and the CARS (e.g., Moulton et al., 2019), have been cited as having substantive correlations between their ASD-related factors. Additionally, prior factor analyses of the GARS/GARS-2 indicated correlated factor solutions (e.g., Lecavalier, 2005; Pandolfi et al., 2010; Volker et al., 2016).

Though information about inter-factor correlations was not reported for the EFA in the GARS-3 manual (Gilliam, 2013), an oblique (promax) rotation was used in the EFA, which suggested that correlations among factors were anticipated. Theoretically, it would be difficult to justify the use of a composite score that subsumes multiple, uncorrelated factor-based subscales. A lack of correlations among factors/subscales would suggest that such factors/subscales each measure something unique, and do not measure something in common with other factors/subscales. As evidenced in the above studies (e.g., Norris et al. 2012) examining inter-factor correlations of ASD symptoms, this is likely not true. Given that each subscale is assumed to assess some specific aspect or a larger construct (i.e., each is a lower-order factor of more general higher-order ASD factor), one would expect lower-order factors to covary.

Research Question 4. When students with ASD are rated by special education teaching staff with the GARS-3, how does the six-factor EFA solution correspond to the six GARS-3 subscales proposed by the author (Gilliam, 2013)?

Rationale. Given the exploratory nature of EFA, there was no specific prediction regarding the number of factors to be retained prior to the analysis. Thus, it is possible that a factor structure different from Gilliam's (2013) proposed structure may be found and retained as most interpretable. Although a six-factor solution may or may not be the retained, most interpretable, factor structure for this EFA study, the available six-factor solution from the EFA was still compared to the author-reported six-factor structure for the sake of thoroughness – to examine the degree of alignment. Overlap between the two solutions was anticipated. Prior research regarding GARS factor analyses have reported the percentage of item agreements between the obtained factors and author-proposed subscales (e.g., Lecavalier, 2005; Pandolfi et al., 2010; Volker et al., 2016). In the majority of prior findings, results retained the same number of factors, but considerable discrepancies were found regarding item placement on factors/subscales (e.g., Lecavalier, 2005; Volker et al., 2016). Therefore, the similarities and differences can be both nuanced and important.

Table 1. Summary of Study One Research Questions

	Research Question	Hypothesis	Analysis	Method
1	How many potentially interpretable factors are present and should be considered for retention?		Scree plot, Kaiser criterion, Velicer's MAP, parallel analysis	EFA with PAF
2	How many factors should be retained to yield the most interpretable factor solution?		Using criteria from the prior analysis to decide the number of potential factor solutions, independent researchers will examine the interpretability of these solutions	EFA interpretive procedure
3	Are there substantive correlations between at least some GARS-3 factors within the most interpretable factor structure?	Correlations between at least some factors will be ≥ .30.	Examine the inter-factor correlation matrix to determine if substantive correlations are present	EFA with oblique rotation
4	How does the six-factor solution correspond to the six GARS-3 subscales proposed by the author (Gilliam, 2013)?		Examine the factor constructs/names of the six-factor solution, as well as item loadings, compared to the six GARS-3 subscales	Qualitative comparison, calculation of the percentage of overlapping items per factor

Note. Velicer's MAP = Velicer's minimum average partial correlation; EFA = Exploratory factor analysis; PAF = Principal axis factoring; GARS-3 = Gilliam Autism Rating Scale – Third Edition

Sample Demographics

The sample's data were collected from a special education agency in Western New York state as part of annual program reviews. The EFA sample consisted of 204 individuals with a diagnosis, or suspected diagnosis, of ASD. The sample was majority male (76.50%) and White (77.50%), with an average age of 9.75 (*SD* = 5.19). The average most recent cognitive assessment results or intelligence quotient (IQ) was 61.48 (*SD* = 21.40) and estimates ranged from 18 to 124. Please refer to a more detailed breakdown of demographic variables below.

Table 2. Study One Demographic Information of Sample

Demographic Variable	
Age in years – *M* (*SD*)	9.75 (5.19)
Most Recent IQ – *M* (*SD*)	61.48 (21.40)
<70 (%)	65.69
≥70 (%)	31.86
Unknown (%)	2.45
Gender (%)	
Male	76.50
Female	23.50
Ethnicity (%)	
Caucasian	77.50
African American	8.30
Latino	8.30
Asian	3.40
Other	2.50
Diagnosis (%)	
ASD	90.20
PDD-NOS	7.35
Suspected Diagnosis	2.45

Note. ASD = Autism spectrum disorder; PDD-NOS = Pervasive developmental disorder, not otherwise specified

It was important to evaluate whether the number of available cases met minimum requirements for conducting the EFA. There are different general recommendations and "rules of thumb" for sample size in the factor-analytic context. Some researchers suggest a sample size of

300 would be ideal, but that a smaller sample of 150-200 would be suitable if the participant to variable ratios were reasonable (e.g., such as 5 participants:1 variable or item) or if there are likely to be high factor loadings (Floyd & Widaman, 1995; Yong & Pearce, 2013). While reasonable, these recommendations are broad generalizations that do not account for other important considerations in making the decision (e.g., amount of common variance likely present in the correlation matrix, level and range of likely item communalities in the diagonal of the correlation matrix, etc.). However, MacCallum et al. (1999) conducted an extensive simulation study to estimate the likelihood of various sample sizes to recover the population factors under varying conditions of item communalities and anticipated item to factor ratio. The authors reported a table (Table 1) that indicated the percentage of successful factor solutions across simulations for various combinations of sample size, approximate ratio of number of factors to number of items, and range of item communalities (MacCallum et al., p. 93). The table is useful for assessing the adequacy of sample sizes for factor recovery, given reasonable estimations of anticipated number of factors, number of items, and range of item communalities. In the GARS-3 manual, the author reported six factors for a scale consisting of 58 items, but item communalities are not reported (Gilliam, 2013). For purposes of item communality estimation, prior GARS-2 factor analyses (e.g., Volker et al., 2016) reported item communalities that fell within MacCallum et al.'s “wide” communality range designation (i.e., communalities that were not generally high or generally low, but were roughly distributed over the .20 to .80 range were considered “wide"). Looking at the options available in the table generated by MacCallum et al., the GARS-3 proposed structure with six factors and 58 items would likely be reflected in the approximate 20:3 (20 items for every 3 factors) ratio – although a more accurate ratio (not represented in the table) would be about 29:3 and the yielded result would likely be better. Using

this information with our sample of 204, close to 100% of the simulated solutions were convergent for any of the available communality designations (i.e., low, wide, or high).

Data Cleaning and Dealing with Missing Data

Data cleaning and preparation followed Osborne and Banjakovic’s (2016) recommendations and procedures. Potential outliers and unusual data values were identified and checked for possible data entry or rating errors and examined for possible undue influence on the analyses. By looking over the dataset carefully and conducting descriptive analyses to identify any values that fell outside the four-point item scale rating, these errors or outliers were addressed by either fixing an error, treating it as missing data, or removing an extreme outlier (However, it was not necessary to remove any outliers in the dataset). Given the quality control procedures followed during data collection, missing data points were expected to be rare. The percentage of missing data very low (i.e., 0.21%) and data met the assumption of missing completely at random. The expectation-maximization algorithm was selected to estimate missing data values because of the advantages of a maximum likelihood procedure and lack of need for a more intensive procedure (e.g., multiple imputation would require averaging over multiple datasets (Dempster et al., 1977; Osborne & Banjanovic, 2016; Roth, 1994). This was computed via the SPSS Missing Values module (IBM Corp., 2019a).

Data Analysis

This analysis involved several programs including SPSS Version 26 (IBM Corp., 2019b), Project R (R Core Team, 2019) ordinal factor analysis options (Basto & Pereira, 2012) accessed via the R-plug-in for SPSS (IBM Corp., 2010), and SAS Version 9.4 (SAS Institute Inc., 2013). SPSS was used for data entry, data storage, data cleaning, and for generating descriptive statistics. Generation of the inter-item polychoric correlation matrix, scree plot, eigenvalues for evaluating the Kaiser criterion, and analyses such as parallel analysis, and Velicer's MAP (Costello & Osborne, 2005; Osborne & Banjanovic, 2016), was accessed through the R ordinal factor analysis menu (Basto & Pereira, 2012) via the R-plug-in for SPSS. The inter-item polychoric correlation matrix was imported into SAS, where the EFA was conducted using PAF as the method of extraction, after evaluation of distribution assumptions.

Input Matrix for the EFA

GARS-3 data are best considered ordinal as item scaling involved the rater assigning one of four discrete ordered categories as a response for each item. As such, an inter-item correlation matrix consisting of polychoric correlations was used for the EFA, as such correlations are better suited to ordinal data than standard Pearson correlations—which assume more continuous, interval-level data. Resulting polychoric coefficients are maximum likelihood estimates of *r* correlations, expected to be equal to or greater than *r* values (Basto & Pereira, 2012).

In order to assess the general suitability of the correlation matrix for factor analysis, the Bartlett's Test of Sphericity (Bartlett, 1950) and the Kaiser-Meyer-Olkin (KMO; Kaiser, 1970) test were used. Bartlett's test examines the basic assumption of whether the correlation matrix is the same or different from an identity matrix. In order to run an EFA, there should be significant correlations in the matrix which would not be present in an identity matrix (Raykov & Marcoulides, 2008). Therefore, a significant Bartlett's test reflects a rejection of the null hypothesis that the correlation matrix is consistent with an expected identity matrix in the presence of sampling error. Furthermore, the KMO is an index indicative of the common variance available in the correlation matrix. The higher KMO value, the more common variance available for estimating common factors in an EFA (Kaiser, 1970). There is a series of standards for KMO, with values ranging from below .50 ("unacceptable") to above .90 ("marvelous"). A value of .60 would be considered “mediocre” while higher values such as .70 (“middling”) and .80 (“meritorious”) are considered more desirable (Kaiser & Rice, 1974, p. 112). The correlation matrix should have, at the very minimum, a KMO above .50 in order to move forward with an EFA (Kaiser & Rice, 1974).

Extraction Methods

The data was assessed for normality assumptions before extraction. Because of normality violations, PAF was used (Costello & Osborne, 2005).

Factor Selection and Retention

To inform factor retention decisions, a number of methods were used to estimate the likely number of interpretable factors present. These methods, as cited as common EFA practices (e.g., Costello & Osborne, 2005), included the Kaiser criterion (Kaiser, 1960), the scree plot (Cattell, 1966), parallel analysis (Horn, 1965), and Velicer’s MAP (Velicer, 1976). The Kaiser criterion and the scree plot are considered more traditional and popular as they are easy to compute with common statistical software programs, however, are considered less accurate strategies (O’Connor, 2000). Additional methods of parallel analysis and Velicer’s MAP are considered more accurate, best practice strategies for determining the number of likely factors present (Basto & Pereira, 2012; O’Connor, 2000; Velicer & Jackson, 1990). The different methods typically suggest a range of different factor solutions to be explored for interpretation and the most interpretable solution will be retained. While each method was explored in the current study, a greater emphasis was placed on the suggestions provided from parallel analysis and Velicer’s MAP.

The Kaiser criterion (Kaiser, 1960) involves retaining all factors that yield an eigenvalue greater than one (Costello & Osborne, 2005). The logic, originally derived from principal components analysis, reflects the belief that a meaningful factor or component should account for more variance than single item or other indicator in the analysis (Braeken & van Assen, 2017). An eigenvalue less than one would indicate factor variance is less than the variance of a single item or indicator (Brown, 2015). The scree plot (Cattell, 1966) was also used to roughly estimate, using the "elbow" or break in the eigenvalue graph, the number of factors to retain (Yong & Pearce, 2013).

For additional and more accurate criterion references, parallel analysis (Horn, 1965) and Velicer's MAP criteria (Velicer, 1976) were used (Costello & Osborne, 2005; O'Connor, 2000). For parallel analysis, eigenvalues of the existing dataset are compared to the eigenvalues generated from random data. Specifically, the literature suggests looking at eigenvalues within the 95th percentile of the random data. Parallel analysis retains factors until the eigenvalue of the random data is greater than the same numbered eigenvalue from the existing dataset. Velicer's MAP criteria retains factors until there is more unsystematic compared to systematic variance. While these two statistical procedures may yield the same number of factors to retain, the literature suggests both should still be used (O'Connor, 2000).

Each of the methods were considered in guiding the number of factors to retain. The different methods varied in terms of the number of factors suggested for retention. Because of this, solutions suggested by parallel analysis and Velicer's MAP were given the most consideration as they are recommended as more accurate strategies (Basto & Pereira, 2012; Costello & Osborne, 2005). For thoroughness, factor solutions were examined that ranged from one less than lowest number of factors suggested, and one more than the highest number of factors suggested. This strategy allowed for a range of possible factor solutions to be explored for interpretability and attempt to account for sampling error and imperfections in the other factor selection methods. Ultimately, the most interpretable factor solution was selected and retained (see interpretation below).

Rotation

An oblique rotation, allowing inter-factor correlations, was used. Because latent constructs reflecting features of ASD are likely to be correlated, a rotation that allows for correlated factors should be initially used (Osborne & Banjanovic, 2016). In the GARS-3 manual, the test author reported using the oblique rotation, promax, though no information regarding obtained inter-factor correlations was reported (Gilliam, 2013). The correlated solution was found to be viable, and the obliquely rotated solution was retained (Nunnally & Bernstein, 1994; Pell et al., 2003).

Interpretation

Information from the extraction, factor retention methods, and rotation provided data for the researcher to interpret different solutions. Typically, researchers should anticipate looking at factor structures with one more and one less factor based on the criterion. Ultimately, the most ideal factor solution would have factor loadings of at least .30, very few – or no – cross loadings, and with each factor containing three or more items (Costello & Osborne, 2005).

Assignment of Items to Factors. In general, an item is assigned to the factor on which it loads highest. This was true for the present study. Items with factor loadings of ≥ .30 are considered substantive (Costello & Osborne, 2005; Yong & Pearce, 2013). If an item loads substantively on more than one factor, this is referred to as a cross-loading. Ideally, each item loads on only one factor (e.g., all items for a factor load only on that factor) resulting in a clearer interpretation of factor scores or factor-based subscales (Costello & Osborne, 2005).

Interpreting and Naming Factors. Factor interpretations and names were assigned by examining the content of the items with the highest loadings on that factor. Typically, items with high loadings on a factor closely align to what the factor is measuring, either directly or indirectly. However, item content across the range of loading values is informative in capturing nuances of the measured construct of the factor. In addition, cross-loading items (though generally undesirable from a scale development perspective) can inform naming procedures. There may be theoretical reasons why a particular item loads on two different factors that have relevant meaning for it (e.g., an item about not listening to instructions might load on both an inattention factor and/or an oppositional behavior factor).

Cross-Validation of Factor Interpretations. For purposes of the present study, five different researchers independently examined the possible factor solutions, interpreted and named the factors, and selected the most interpretable/meaningful factor solution. These evaluators were all ASD researchers – advanced doctoral students and faculty – familiar with the core and associated features of ASD and other developmental disorders. The generated factor names were examined for conceptual similarity and independently chosen solutions were examined for convergence. Any discrepancies were resolved through discussion and final consensus. Prior exploratory factor analyses (e.g., Nelson et al. [2016] with the SRS-2; Volker et al. [2016] with the GARS-2) have used this peer-review method.

Content/Item Comparison

The six-factor solution was qualitatively compared, at the construct level, to the six subscales in the GARS-3 proposed by the author (Gilliam, 2013). Factors and existing subscales were compared in terms of name and construct similarity—taking factor loadings and general item content into account. The percentage of overlapping items was calculated for each factor with the subscale that is its closest construct match (i.e., number and percentage of items that overlap on similar factors between the six-factor EFA solution and the GARS-3 author-proposed six-factor model).

Internal Consistency

Internal consistency reliability is reported for the study one factor-based subscales. Estimates are reported as both Cronbach's coefficient alpha (Cronbach, 1951) and ordinal alpha (Zumbo et al., 2007). The ordinal alpha provides an internal consistency estimate that does not assume continuous variables and corrects the alpha coefficient for the ordinal nature of the rating scale data (Zumbo et al., 2007). Though the ordinal alpha may fit the data model better, reporting the Cronbach's alpha allows for two important comparisons. First, comparing the Cronbach's and ordinal alpha values will give a sense of the degree to which the ordinal correction changed the alpha value. Second, because the GARS-3 manual (Gilliam, 2013) reported Cronbach's alpha values based on the standardization sample, reporting them for the present study will allow for a more direct comparison.

Study Two: CFA

Research Questions, Rationale, and Hypotheses

Research Question 5. When students with ASD and non-ASD developmental disabilities (DDs) are rated by special education teaching staff with the GARS-3, does the interpretive model proposed by the GARS-3 test author produce a reasonable fit to the confirmatory sample covariance matrix?

Hypothesis 5. It was predicted that the GARS-3 author's proposed six-factor solution will yield a reasonable fit to the confirmatory sample inter-item covariance matrix.

Rationale. The model proposed for this prediction is a statistical representation of what would be the expected latent variables based on the published GARS-3 subscales and scoring. As reported in the manual, the author found a six-factor structure through a prior EFA using the standardization sample which is consistent with the six proposed subscales for the GARS-3 (Gilliam, 2013). As noted in the EFA section, there are sound reasons to believe and assume these subscales are inter-correlated because the GARS-3 subscales proposed in the manual represent constructs consistent with the presence of ASD and because there is a composite score. However, the test author does not report inter-factor correlations, nor were any fit indices reported despite the use of an EFA factor extraction procedure (ML) that would yield indices of fit (Gilliam, 2013). Please refer to Appendix B for a visual depiction of the correlated six-factor structure. In this diagram, the circles represent each factor, the squares represent items assigned to each factor, and the ovals indicate the items' error/disturbance terms.

Research Question 6. When students with ASD and non-ASD DDs are rated by special education teaching staff with the GARS-3, does the retained factor solution from the study one EFA produce a reasonable fit to the confirmatory sample inter-item covariance matrix?

Hypothesis 6. It was predicted that the retained factor solution from study one will reasonably fit the inter-item covariance matrix from the confirmatory sample.

Rationale. It was hypothesized that the factor solution would reasonably fit the confirmatory sample for a number of reasons. First, the model from study one was developed using best practice EFA methodology (e.g., polychoric correlation matrix for ordinal data, use of robust factoring procedure for non-normal data, use of parallel analysis and Velicer's MAP as part of the factor selection process; Basto & Pereira, 2012; Costello & Osborne, 2005; O'Connor, 2000). Further, the CFA sample was similar to the EFA sample from which the factor model was generated, and the same rater types were used. Given the non-normal item data, PAF was needed for factor extraction in the study one EFA, which did not provide indices of fit. To accommodate predicted ordinal and non-normal data, the CFA used a diagonally weighted least squares (WLS) factor extraction procedure – highly robust to data violations and yields appropriately adjusted fit indices (e.g., the weighted least squares mean variance procedure [WLSMV] in Mplus [Brown, 2015; Muthen & Muthen, 1998-2017]). Therefore, the purpose of the CFA in this study was to cross-validate the retained EFA solution, assess its generalizability outside of the original EFA sample, and provide assessment of model fit to the covariance matrix using robust methods (Brown, 2015). No previous work of this kind has been reported in the literature for the GARS-3 (i.e., no prior independent EFA and CFA).

Research Question 7. When students with ASD and non-ASD DDs are rated by special education teaching staff with the GARS-3, and the GARS-3 author-proposed model and the EFA-generated model from study one are compared, does one model show evidence of better fit to the confirmatory sample inter-item covariance matrix?

Hypothesis 7. When compared to the author-proposed model, it was predicted that the EFA-generated model from study one will show a substantively better fit with the inter-item covariance matrix from the confirmatory sample.

Rationale. From a broad research question perspective, it is important for both theoretical and practical purposes to know which of the available factor models for an instrument fit better than others. Ideally, one model stands out in terms of comparative fit and would be assessed for generalization across different populations – as generalizability to different sample variations is an important aspect of factor analysis (Brown, 2015; Floyd & Widaman, 1995). However, it was also specifically predicted that the factor model retained from the study one EFA would fit the confirmatory sample covariance matrix better than the author-proposed model (Gilliam, 2013). Two major reasons for this prediction include the best practice approach to the EFA in study one and also in the similarity of the sample across studies one and two. The methodology of the EFA was more thorough and consistent with best practices as compared to the EFA reported in the GARS-3 manual. It is possible that the EFA reported in the manual may have missed a better fitting solution. Second, the sample characterization from factor analyses of study one and two – in both the individuals being rated and those completing ratings – are different from those reported in standardization sample used for the EFA in the GARS-3 manual (Gilliam, 2013). Therefore, any potential differences in the inter-item covariance matrix related to differences in sample characteristics or rater types would likely better fit the model from study one EFA.

Table 3. Summary of Study Two Research Questions

	Research Question	Hypothesis	Analysis	Method
5	Does the interpretive model proposed by the GARS-3 test author produce a reasonable fit to the confirmatory sample inter-item covariance matrix?	The GARS-3 author's proposed six-factor solution will yield a reasonable fit to the confirmatory sample inter-item covariance matrix.	χ^2, SRMR, RMSEA, CFI, TLI	CFA with WLSMV
6	Does the retained factor solution from the study one EFA produce a reasonable fit to the confirmatory sample inter-item covariance matrix?	The retained factor solution from study one will reasonably fit the confirmatory sample inter-item covariance matrix.	χ^2, SRMR, RMSEA, CFI, TLI	CFA with WLSMV

Table 3 (cont'd)

7	When the GARS-3 author-proposed model and the EFA-generated model from study one are compared, does one model show evidence of better fit to the confirmatory sample inter-item covariance matrix?	The EFA-generated model from study one will show a substantively better fit with the inter-item covariance matrix of the confirmatory sample.	Mplus DIFFTEST (adjusted χ^2) and/or AIC and BIC	CFA with WLSMV, MLR

Note. GARS-3 = Gilliam Autism Rating Scale – Third Edition; SRMR = Standardized root mean square residual; RMSEA = Root mean square error of approximation; CFI = Comparative fit index; TLI = Tucker-Lewis index; CFA = Confirmatory factor analysis; WLSMV = Weighted least squares mean variance; EFA = Exploratory factor analysis; AIC = Akaike information criterion; BIC = Bayesian information criterion; MLR = Robust maximum likelihood.

Sample Demographics

Data, as mentioned above, were collected from annual agency-wide program reviews within a special education agency in Western New York state. The CFA sample consisted of 200 individuals. Similar to study one, this sample was majority male (75.50%) and White (72.00%). The mean age was 8.85 (*SD* = 4.78) with most recent assessment of cognition (i.e., IQ) ranging from 27 to 120 (*M* = 61.50, *SD* = 20.87). The sample majority was comprised of individuals with ASD diagnoses (68.50%); a further breakdown of diagnoses is present in the following paragraph and table.

Table 4. Study Two Demographic Information of Sample

Demographic Variable	
Age in years – *M* (*SD*)	8.85 (4.78)
Most Recent IQ – *M* (*SD*)	61.50 (20.87)
<70 (%)	60.00
≥70 (%)	32.50
Unknown (%)	7.50
Gender (%)	
Male	75.50
Female	24.50
Ethnicity (%)	
Caucasian	72.00
African American	10.50
Latino	11.00
Asian	2.50
Native American/Pacific Islander	0.50
Other	3.00
Diagnosis (%)	
ASD	68.50
No ASD Diagnosis	31.50

Note. ASD = Autism spectrum disorder

The validation sample for the CFA consisted of item data collected in different years from those used in the EFA sample. The cases for the CFA will be a combination of completely new cases and some individuals will overlap with the EFA but will have been rated by a different special education staff member in a different program evaluation year. This sample includes individuals with ASD and non-ASD DDs. This type of sample likely reflects the range of case types that would be present in the population of individuals assessed using the GARS-3 as a level two screener (i.e., broader developmental disabilities cases suspected of ASD). Table 5, below, indicates the non-ASD diagnoses of the 63 individuals in study two. Importantly, one individual may have multiple diagnoses that contribute to the makeup of the sample.

Table 5. Study Two Non-ASD (n = 63) Diagnostic Information of Sample

Diagnosis (%)	
Language Disorder	53.97
Suspected ASD	15.87
Fragile X Syndrome	4.76
Attention Deficit/Hyperactivity Disorder (ADHD)	1.59
Epilepsy	1.59
Oppositional Defiant Disorder (ODD)	1.59
Developmental Delay	1.59
Psychosis	1.59
Tuberous Sclerosis	1.59
Other	14.29
No Diagnosis	25.40

Note. Some cases had multiple comorbid diagnoses and therefore, the percentages column does not equal 100%; ASD = Autism spectrum disorder.

Data Cleaning and Missing Data

The same procedures listed above, under study one, regarding data cleaning (i.e., as referenced in Osborne and Banjakovic [2016]) and missing data were used in study two. This included expectation-maximization to impute missing values (Dempster et al., 1977) via the SPSS Missing Values module (IBM Corp., 2019a).

Data Analysis

In this analysis, two different models were examined via confirmatory factor analytic procedures: the six-factor structure of the published GARS-3 (Gilliam, 2013) and the factor structure yielded from study one. As with study one, multiple statistical software packages were used in this study including SPSS Version 26 (IBM Corp., 2019b) for generating sample and item descriptive statistics and Mplus Version 8.2 (Muthén & Muthén, 1998-2017) for conducting the actual CFAs.

Input Data for the CFA

Similar “rules of thumb” regarding ideal sample size of an EFA apply to CFA. These include at least 100-200 cases and participant-to-item ratio of 10:1 or 5:1. When considering how the sample size might affect statistical power, one type of method used is a Monte Carlo approach which accounts for model parameters such as model, sample size and data (Brown, 2015). The Monte Carlo approach was used in the simulation study that MacCallum et al. (1999) conducted, which was referenced in relation to estimating the EFA sample size from study one. Similar to the EFA sample, a sample size of 200 cases were available for the CFA. As before, given a conservative item to factor ratio of 20:3, close to 100% of simulations yielded a convergent solution with factor recovery assuming any general range of item communality estimates (i.e., low, high, or wide).

Before proceeding with data analysis, data was assessed for univariate and multivariate normality in addition to general suitability for CFA. Based on prior research with the GARS-2 in samples from this special education agency, it was anticipated that the data would not meet normality assumptions. Data did not meet normality assumptions and use of conventional CFA procedures like ML would result in biased fit indices (Brown, 2015). Consistent with the EFA from study one, a polychoric correlation matrix will be used as input for the CFA given the ordinal nature of the data (Basto & Pereira, 2012). Given that the data were both ordinal and non-normal, the most robust available approach was used: a diagonally weighted least squares (WLS) estimation procedure (DiStefano & Morgan, 2014), such as the diagonally WLS mean variance (WLSMV) estimator available in Mplus (Brown, 2015; Muthen & Muthen, 1998-2017).

Model Specification and Identification

As noted, at least some level of basic theoretical and/or empirical basis is assumed prior to conducting a CFA. From this base of support, researchers specify a model (Brown, 2015). In the hypothesized model, researchers specify the variables – both observed and latent – and the number and types of relationships between these variables (Brown, 2015; Byrne, 2012). The broad topic of model identification is complex and nuanced, but in the context of CFA, the two most critical features of model identification are establishing the degrees of freedom (*df*) for the model to be tested and to provide scaling for the latent variables in the model (Byrne, 2012). In terms of considerations for *df*, models fall into three broad categories. These categories are: (a) under-identified models, (b) just identified models, or (c) over-identified models. These designations are based only on the model's *df*, which in the CFA context is the difference between the number of information components available in the variance-covariance matrix of the observed variables and the number of model parameters to be estimated (Brown, 2015). An under-identified model means that there are more parameters to be estimated in the model than there are available pieces of observed information in the variance-covariance matrix ($df < 0$). In the case of a just identified model, the number of model parameters to be estimated equals the number of observed pieces of information in the input matrix ($df = 0$). An over-identified model indicates that there are more available observed pieces of information than model parameters to be estimated ($df > 0$). Critically, an under-identified model cannot be uniquely estimated or tested, a just-identified model can be estimated but cannot be tested (i.e., cannot be subject to possible rejection), and an over-identified model can be both estimated and tested. Thus, an over-identified model is essential for model testing in CFA (Byrne, 2012).

The second critical aspect of model identification involves scaling the latent variables. Latent variables are unobserved and, as a result, have no inherent scale of their own for relating to other variables in the model (Brown, 2015; Byrne, 2012). The two most common approaches to latent variable scaling are the reference variable method (e.g., fixing the factor loading of one observed variable that loads on a factor to 1.0) and the fixed factor method (e.g., setting the factor variance to a specific value, usually 1.0; Byrne, 2012). By fixing the factor loading of one of the observed variables that loads on a factor, the latent variable, in a sense, borrows that variable's scale. This is the most popular method of scaling a latent variable. However, a consequence of this method, is that the fixed factor loading will no longer be freely estimated in the model. With the fixed factor method, wherein the variance of the latent variable is fixed to 1.0, then all factor loadings associated with that factor can be freely estimated. But, by fixing the factor variance, that variance will no longer be freely estimated in the model.

In the present study, two different GARS-3 measurement models were examined. In CFA, models have fixed and possibly estimated parameters such as factor loadings, and unique (i.e., error) and factor variances (Brown, 2015). The first model tested is that described in the GARS-3 manual, consisting of six first-order factors (reflecting six subscales), and involving 58 items. Though the test author did not report inter-factor correlation results for the EFA described in the test manual, inter-factor correlations were clearly anticipated – based on the rotation used – and first-order factor model relationships implied by use of a composite score for a measure. In a more complex, higher-order model, this composite (e.g., Autism Index 6) would theoretically reflect an assumed second-order factor that explains inter-factor correlations (Brown, 2015). However, only the first-order factor models were examined in hypotheses for research questions five through seven, which are prerequisites for examining a more complex, higher-order model. As indicated above, in identification of the model, the latent variables or factors must be assigned an identified scale since they are unobserved variables (e.g., fixing with the factor loadings or the factor variance to 1.0; Byrne, 2012). In the present study, factor loadings were freely estimated for all items, while estimation of exact factor variances was not essential. Thus, the factor variances were fixed to 1.0 to provide scaling for each factor. Importantly, the amount of available information regarding the observed variables (e.g., number of correlations, number of variances, etc.) exceeded the number of freely estimated parameters in the proposed model required to both estimate and test the model (i.e., an over-identified model where *df* > 0; Brown, 2015). Given the number of items involved and the known model, the model was over-identified.

The second model resulted from the EFA of study one. Model specifications of parameters are reported below and, as with the previous model, included factor loadings, unique variances, and factor variances. Additionally, latent variables were scaled such that factor variances were fixed to 1.0 and all factor loadings were freely estimated. Further, the *df* of the model was assessed and the model was over-identified.

Model Estimation and Fit

For model estimation, Mplus Version 8.2 (Muthén & Muthén, 1998-2017) was used. As in the previous study, the item data was explicitly recognized as ordinal and the input matrix consisted of polychoric correlations (Basto & Pereira, 2012).

To conduct a CFA, where parameters are estimated and the model is tested, the pre-specified model (e.g., latent and observed variables, factor loadings, factor variances and covariances, and unique or error variances and covariances) must be appropriately constrained, and over-identified. The model fitting function, also called an estimator or estimation procedure, attempts to achieve the smallest differences possible between the predicted covariance matrix, based on the model, and the actual covariance matrix. Although ML is the most widely used method for this purpose, its strong assumptions of multivariate normality and continuous data was not met by the GARS-3 item data. Thus, a robust diagonally weighted estimator (i.e., Mplus' Weighted Least Squares Mean Variance (WLSMV) estimator), was used (Brown, 2015).

Fit, including how well or how strained the model fits to the data, was assessed in several ways. Different types of fit indices may convey the absolute fit of the model (e.g., fit relative to a hypothetical model where the predicted and actual covariance matrices are equal), involve a parsimony correction (i.e., involve a penalty for more complex models), may convey information regarding comparative/incremental fit, or may be scaled in terms of lack of fit (i.e., scaled to index residual variance unaccounted for, as opposed to the overlap or variance accounted for). To evaluate the model's absolute fit with the data, a chi-squared (χ^2) test was used and the *df* and *p* values are reported. If this test is statistically significant, it indicates lack of fit between the predicted model and the actual data. However, the χ^2 test is not often used in isolation to measure goodness of fit, because statistical significance may be affected by underlying distributions, sensitivity to sample size (e.g., high statistical power in larger samples can render even non-meaningful differences statistically significant), and/or its stringent perfect-fit null standard (Brown, 2015). Additionally, the standardized root mean square residual (SRMR) was used as another indicator of absolute fit to examine the average differences between correlations in the predicted and actual matrix. With this indicator, ranging from zero (i.e., perfect fit) to one, a smaller number is indicative of a better fitting model (Brown, 2015). Hu & Bentler (1999) suggested SRMR values should be equal to or less than .08 to indicate a "reasonably good fit" between the model and data (Brown, 2015).

Further, to evaluate fit from a parsimony-corrected perspective, the root mean square error of approximation (RMSEA; Steiger & Lind, 1980), the Akaike information criterion (AIC; Akaike, 1987), and the Bayesian information criterion (BIC; Schwarz, 1978) were used. Parsimony-corrected fit indices examine model fit but take into account, and penalize the model, for having more freely estimated parameters as indicated by the *df*. RMSEA is used and recommended by researchers to examine relative fit, compared to the more absolute fit mentioned above. Scores typically range from zero (i.e., perfect fit) to one, with smaller values indicating a better fit (Brown, 2015). Hu & Bentler (1999) suggested that the RMSEA should be about equal to or less than .06. Further, Browne and Cudeck (1993) suggested that a value below .05 indicates good fit – rather than adequate values of less than .08 – and rejecting models with values greater than or equal to .10. The AIC and BIC were also used to indicate fit and are specific to models that have non-nested data, with the BIC being more critical of freely estimated parameters (i.e., BIC involves a more stringent parsimony correction that penalizes more complex models). It should be noted that these are not inferential statistical calculations, but they are commonly used descriptive indicators for comparing the fit of alternative or competing models. When models are compared, lower AIC and BIC values are indicative of an overall better fit (Brown, 2015). Thus, the model with the lower AIC or BIC value would be selected. (Note that the AIC and BIC indices are not available when the WLSMV estimator is used [Muthen & Muthen 1998-2017]. Since the WLSMV was required, due to data conditions, WLSMV-generated fit indices were supplemented with AIC and BIC values generated using the robust maximum likelihood estimator (MLR). In this situation, the AIC and BIC were used only for comparing models and were not used to make individual model decisions with the other indices generated by WLSMV.)

Within the last category of fit indices, comparative fit indices such as the comparative fit index (CFI; Bentler, 1990) and the Tucker-Lewis index (TLI; Tucker & Lewis, 1973) were examined. A comparative fit index seeks to compare the predicted model to a restricted, nested model, typically with covariances set to zero (Brown, 2015). The CFI uses the χ^2 and *df* of the predicted model and comparison null baseline model, such that resulting values closer to one indicate better fit. Similar to the CFI, the TLI is calculated using the χ^2 and df with higher values indicative of better fit. Like the RMSEA, the TLI penalizes models with more freely estimated parameters. Though scaled and interpreted similarly to the CFI, because the TLI is not normed, the resulting TLI values can exceed 1.0 (Brown, 2015). Hu & Bentler (1999) suggested that both CFI and TFI values should be about .95 or greater (Brown, 2015).

To directly compare the fit of the GARS-3 author-proposed model and the EFA-generated model from study one, the Mplus DIFFTEST and AIC and BIC indices were intended to be used (Brown, 2015; Muthén & Muthén, 1998-2017). The Mplus DIFFTEST can only be utilized when one of the models is nested within the other (Muthén & Muthén, 1998-2017). Additionally, AIC and BIC indices were used descriptively for cross-model comparisons (Muthén & Muthén, 1998-2017). As mentioned above, the AIC and BIC are not available through the WLSMV estimator and were obtained using supplemental MLR estimations for the two models (see Birnbaum, 2020).

Study Three: Clinical Discriminant Validity

Research Questions, Hypotheses, and Rationale

Research Question 8. When students with developmental disabilities, in a center-based special education setting, are rated by special education teaching staff using the GARS-3, how well does the GARS-3 discriminate individuals with ASD from individuals with other developmental disabilities that require substantial support?

Hypothesis 8a. It is predicted that the mean GARS-3 Autism Index 6 score for students with ASD will be significantly higher than the mean Autism Index 6 score for students with other developmental disabilities.

Hypothesis 8b. Using the author-recommended cut score of 70 on the Autism Index 6, the sensitivity for accurately identifying risk level for cases with ASD will be ≥ .90.

Hypothesis 8c. Using the author-recommended cut score of 70 on the Autism Index 6, the specificity for accurately identifying those not at risk for ASD will be ≥ .80.

Exploratory analysis (8d). Determination of the optimal cut score for the GARS-3 Autism Index 6. This exploratory analysis will be conducted using Receiver Operating Characteristic (ROC) Curve analysis to examine the range of possible cut scores on the GARS-3 Autism Index 6 for purposes of determining the optimal screening cut score in the context of study three sample.

Rationale. The purpose of these analyses was to examine how well the GARS-3 can discriminate ASD from other developmental disorders that require significant support (e.g., ID). Theoretically, a scale that measures ASD, and associated constructs, should have a higher score for students with confirmed ASD diagnoses and thus, should be able to differentiate between known groups that differ on this construct. The GARS-3 manual (Gilliam, 2013) included reports of relevant mean differences, sensitivity, specificity, and ROC/Area Under the Curve (AUC) estimates. The mean differences reported in the manual indicated that the ASD sample's Autism Index scores were significantly greater than the comparison groups (i.e., ID, ADHD, ED/BD, LD, SLI, TD; Gilliam, 2013), a result that is consistent with studies involving the GARS-2 composite score (e.g., see Volker et al., 2016). With GARS-3, which has been through multiple editions, it is likely more important to examine the size of the difference between the different groups and not whether there is any difference at all in the expected direction.

Based on the recommended GARS-3 cut score of 70, the sensitivity and specificity values reported in the manual were quite high (i.e., .96 and .84, respectively) for differentiating between ASD and a group consisting of a number of other disabilities (i.e., ID, deaf, blind, ADHD, ED/BD, LD, and physical/health impairment; Gilliam, 2013). However, sensitivity and specificity values reported in manuals for prior versions of the GARS, appeared to be superior relative to values reported by independent researchers using samples recruited through registry, agency, educational, and clinical settings (e.g., Lecavalier, 2005; South et al., 2002; Volker et al., 2016). This was especially true for sensitivity values, which were considerably lower in the external samples. With the GARS-3 (Gilliam, 2013), the author may have attempted to account for these findings as there was a new, lower recommended a lower cut score compared to prior editions (i.e., Autism Index > 70 for GARS-3 [Gilliam, 2013], ≥ 85 for GARS-2 [Gilliam, 2006], and ≥ 90 for GARS [Gilliam, 1995]). Of further importance, 42 of the 58 items of the GARS-3 are new relative to prior editions. This may or may not impact the sensitivity and specificity, however, the general need for cross-validation and the discrepancies noted for prior editions of the instrument clearly support the need to independently examine the sensitivity and specificity of the instrument.

Because of the difficulty in discriminating between ASD and other developmental disorders, which share some similar characteristics, it is important to examine if a different cut score could be used to improve discriminant validity with ROC curve analysis. The ROC/AUC reported in the manual, comparing the same ASD and not ASD disabled sample – including those with ID – was .89 for the Autism Index 6. Although the manual indicated significant differences between the ASD normative sample and other diagnostic groups, the mean Autism Index scores of the ID sample (*n* = 15) were well above the cut score of 70 (i.e., Autism Index 6 *M* = 87, *SD* = 22; Gilliam, 2013). Research with other ASD scales have acknowledged the use of different cut scores depending on the sample (e.g., using the SCQ to discriminate ASD from other developmental disabilities; Wiggins et al., 2007). There is clearly a need for further examination of the cut score with ASD and other developmental disabilities involving substantial impairment. The rationale for the ROC curve analyses is derived from the need to assess the generalizability of the recommended cut score in other samples or involving other discriminant situations, different rater types, etc. The ROC analyses allowed for examination of all possible cut scores with their associated sensitivities and specificities.

Table 6. Summary of Study Three Research Question

	Research Question		Hypothesis	Analysis/Method
8	How well does the GARS-3 discriminate between individuals with ASD from individuals with other developmental disabilities that require substantial support?	8a	The mean GARS-3 Autism Index 6 score for students with ASD will be significantly higher than the mean Autism Index 6 score for students with other developmental disabilities.	Mean differences, independent samples *t*-test, Cohen's (1988) effect size *d*
		8b	The sensitivity for accurately identifying risk level for cases with ASD will be ≥ .90.	Sensitivity calculation
		8c	The specificity for accurately identifying those not at risk for ASD will be ≥ .80.	Specificity calculation
		8d	*Exploratory analysis*: Determination of the optimal cut score for the GARS-3 Autism Index 6.	ROC Curve analysis to examine the range of potential cut scores to determine the optimal cut score for the study three samples

Note. GARS-3 = Gilliam Autism Rating Scale – Third Edition; ROC = Receiver operating characteristic.

Sample Demographics

The study three sample consisted of two groups. The first group consisted of the same ASD sample as study one (*n* = 204), plus the unique cases from study two (*n* = 22), totaling 226 individuals in the ASD group. The second group consisted of cases with non-ASD developmental disabilities, including ID, that require substantial support (*n* = 64). Like the ASD sample, the non-ASD sample were also students who receive center-based special education services and attend the same special education agency. This group is also characterized by similar levels of functional impairment including cognitive scores, adaptive behavior, language and motor skills, and educational and behavioral needs. Specifically for most recent IQ information, the ASD sample ranged from 18 to 124 (i.e., the same as study one) and the non-ASD sample most recent IQ data ranged from 30 to 113. Table 7 depicts the demographic information of the combined ASD and non-ASD sample, in addition to demographics separate to each diagnostic group. Table 8 provides information for the specific diagnoses of the non-ASD sample. The compilation of this data includes all diagnoses, including comorbidities (e.g., one individual may have multiple diagnoses that contribute to the makeup of the sample).

Table 7. Study Three Demographic Information of Sample

Demographic Variable	Combined (*n* = 290)	ASD (*n* = 226)	Non-ASD (*n* = 64)
Age in years – *M* (*SD*)	9.20 (5.25)	9.79 (5.16)	7.12 (5.04)
Most Recent IQ – *M* (*SD*)	62.61 (21.90)	60.78 (21.23)	70.88 (23.16)
<70 (%)	59.31	66.81	32.81
≥70 (%)	33.79	30.97	43.75
Unknown (%)	6.90	2.21	23.44
Gender (%)			
Male	74.80	74.30	76.60
Female	25.20	25.70	23.40
Ethnicity (%)			
Caucasian	73.80	77.00	62.50
African American	9.70	9.30	10.90
Latino	10.00	8.00	17.20
Asian	3.10	3.10	3.10
Native American/Pacific Islander	0.30	0.00	1.60
Other	2.80	2.70	3.10
Diagnosis (%)			
ASD	77.90	100.00	0.00
No ASD Diagnosis	22.10	0.00	100.00

Note. ASD = Autism spectrum disorder; PDD-NOS = Pervasive developmental disorder, not otherwise specified

Table 8. Study Three Non-ASD (n = 64) Diagnostic Information of Sample

Diagnosis (%)	
Language Disorder	53.13
Suspected ASD	14.06
Fragile X Syndrome	4.69
ADHD	1.56
Epilepsy	1.56
ODD	1.56
Developmental Delay	1.56
Psychosis	1.56
Tuberous Sclerosis	1.56
Other	14.06
No Diagnosis	26.56

Note. Some cases had multiple comorbid diagnoses and therefore, the percentages column does not equal 100%.

Data Cleaning and Missing Data

The same protocols for data cleaning and missing data from study one and study two were used in study three. This included recommendations from Osborne and Banjakovic (2016) and use of estimation-maximization to impute missing values (Dempster et al., 1977).

Data Analysis

To analyze clinical discriminant validity, three methods were used: mean difference comparisons, sensitivity/specificity of pre-determined cut score, and exploratory ROC curve (Metz, 1978) analyses. Mean differences were examined between Autism Index scores. The mean Autism Index scores from the ASD group and the group with other developmental disabilities were compared using an independent samples *t*-test at an alpha level of .01. Because the Autism Index score of the ASD sample was predicted as being higher than the sample with other developmental disabilities, a one-tailed test was used. The effect size was calculated using Cohen's *d* (Cohen, 1988). These analyses were computed through SPSS Version 26 (IBM Corp., 2019b).

Using the cut score of 70 (i.e., 71 or higher = “very likely” probability of ASD, according to the GARS-3 record form and manual; Gilliam, 2013) and knowledge of the sample’s previously established diagnoses, the researcher created a 2 x 2 table with the number and percentage of cases identified as true positives, true negatives, false positives, and false negatives. Sensitivity and specificity were calculated based on these numbers. Sensitivity was calculated by dividing the number of true positives by the sum of the true positives and false negatives. Specificity was calculated by dividing the number of cases deemed as true negatives by the sum of true negatives and false positives. Please refer to Table 22 in the results section for a model of a binary classification matrix. Additionally, positive predictive power was calculated by dividing the number of true positives by the sum of true positives and false positives, and negative predicted power was calculated by dividing true negatives by the sum of true negatives and false negatives (Bandalos, 2018; Lalkhen & McCluskey, 2008).

Further, ROC curve analyses (Metz, 1978) were used to examine all possible screening cut scores and associated sensitivity and specificity values in order to determine the optimal cut score for this specific sample. This analysis was intended to assess how well the pre-determined cut score generalizes to the current study sample and to possibly suggest a replacement of that cut score when discriminating between similar samples. Through ROC curve analyses, the cut score can be adjusted (i.e., raised or lowered) thereby changing the sensitivity and specificity. When the cut score is lowered, in this case below 70, there will be likely higher rates of false positives as more individuals would be considered likely to have ASD. When the cut score is raised, above 70 for the GARS-3, there may be more false negatives or individuals falsely identified as being likely to have ASD (Lalkhen & McCluskey, 2008). Through this process, the researcher sought to balance the sensitivity and specificity of the instrument with an optimal cut score for this specific sample that includes both ASD and non-ASD developmental disorders. These analyses were conducted through SPSS Version 26 (IBM Corp., 2019b).

CHAPTER FOUR: RESULTS

Study one examined the factor structure of the GARS-3 when students with ASD (*n* = 204) were rated by special education teaching staff. Specifically, this study sought to determine the number of potentially interpretable factors for retention, to examine the presence of substantive correlations between the factors, and to assess the overlap, in terms of overlapping factors and factor item content, between the factor model from the study one exploratory factor analysis (EFA) and the six-factor model proposed by the test author (Gilliam, 2013). To answer these research questions, an EFA was conducted using the inter-item polychoric correlation matrix as input. Study two utilized confirmatory factor analysis (CFA) to examine the fit of both the model from study one and the published six-factor model with a sample of students with ASD and non-ASD developmental disabilities (DDs; *n* = 200) rated on the GARS-3 by special education teaching staff. This particular study also sought to examine if one of these specified models demonstrated evidence of a better fit in the confirmatory sample. Lastly, study three assessed clinical discriminant validity of the GARS-3. The study identified how well the measure could discriminate individuals with ASD from individuals with other developmental disabilities that require substantial support (*n* = 290). Analyses for study three involved comparison of mean differences for subscale and composite scores between the two discriminant groups, sensitivity and specificity, and Receiver Operating Characteristic (ROC) Curve analysis to examine possible cut scores for the GARS-3 Autism Index 6. Results are presented for each of the three studies in relation to study-specific subsets of the eight research questions that guided this dissertation.

Data Cleaning and Missing Data

Frequencies were run for items to assess if any values fell outside the appropriate scale range and to initially identify missing values. The researcher assessed the IDs across the rows of the dataset to ensure there were no redundancies in cases (i.e., each case occurs only once in the dataset). Once any values that fell outside the appropriate range were corrected and missing values were verified as truly missing (verified using the original record form from which data were collected), data were then assessed for the percentage of missing values. The data were evaluated to have 0.21% missing items across cases, equal to 25 total missing item values out of 11,832 item scores (i.e., 204 cases x 58 GARS-3 items per case). For those cases with missing values, missing items ranged from one to five per case, with the mode being one missing item per case. Expectation-maximization (Dempster et al., 1977) was used to generate values for the missing items.

Study One: EFA

Data Matrix Sufficiency for Factoring

The table below (Table 9) presents the descriptive statistics for each of the 58 items of the GARS-3 from the study one dataset. Items reported in all tables of this project are truncated (see Appendix F). The mean and standard deviation of each item is presented. Additionally, the percent of responses per item is presented. (Each item was rated on a four-point scale ranging from 0 ["not at all like the individual"] to 3 ["very much like the individual"; Gilliam, 2013].)

Table 9. Study One Dataset Item-Level Descriptive Statistics

Item	Item Stem	Mean	Standard Deviation	Percent of Responses Per Item			
				0	1	2	3
				Not at all	Not Much	Some-what	Very much
				like the individual			
1	Majority of time alone spent in repetitive or stereotyped behaviors	1.735	1.148	21.1	19.1	25.0	34.8
2	Preoccupied with specific stimuli	1.407	1.112	27.9	25.0	25.5	21.6
3	Stares at hands, objects, or items in environment	1.328	1.147	34.8	17.2	28.4	19.6

Table 9 (cont'd)

4	Flicks fingers rapidly in front of eyes	0.672	0.990	62.7	15.2	14.2	7.8
5	Makes rapid lunging, darting movements	0.941	1.135	52.5	15.7	17.2	14.7
6	Flap hands or fingers	0.828	1.057	55.4	16.7	17.6	10.3
7	Makes high-pitched sounds or other vocalizations	1.245	1.255	43.1	15.2	15.7	26.0
8	Uses toys or objects inappropriately	1.103	1.129	44.1	16.2	25.0	14.7
9	Does things repetitively	1.480	1.176	29.9	18.6	25.0	26.5
10	Engages in stereotyped behaviors in play	1.466	1.193	32.4	14.7	27.0	26.0
11	Repeats unintelligible sounds	1.319	1.192	37.3	16.2	24.0	22.5
12	Shows unusual interest in sensory aspects	1.093	1.126	42.6	21.6	19.6	16.2
13	Ritualistic or compulsive behaviors	1.431	1.136	28.4	23.5	24.5	23.5
14	Does not initiate conversations	1.966	1.151	18.6	11.8	24.0	45.6
15	Pays little or no attention to peers	1.613	1.097	21.6	22.1	29.9	26.5
16	Fails to imitate	1.461	1.057	22.1	30.9	26.0	21.1
17	Doesn't follow other's gestures to look at something	1.427	0.997	21.6	29.9	32.8	15.7
18	Seems indifferent to other person's attention	1.578	1.118	24.0	20.1	29.9	26.0
19	Shows minimal expressed pleasure in interactions	1.382	1.018	24.5	27.9	32.4	15.2
20	Displays little or no excitement in showing toys or objects	1.672	1.151	22.1	21.6	23.5	32.8
21	Seems uninterested in pointing out things	1.765	1.197	23.0	16.7	21.1	39.2
22	Seems unwilling to get others to interact	1.574	1.174	25.0	24.0	19.6	31.4
23	Shows minimal or no response when others attempt to interact	1.382	1.060	26.0	27.9	27.9	18.1
24	Displays little or no reciprocal communication	1.402	1.099	26.0	29.9	22.1	22.1
25	Doesn't try to make friends	1.848	1.208	21.6	16.7	17.2	44.6

Table 9 (cont'd)

26	Fails to engage in creative play	1.686	1.195	24.5	18.1	21.6	35.8
27	Shows little or no interest in others	1.304	1.039	28.9	26.0	30.9	14.2
28	Responds inappropriately to humorous stimuli	1.637	1.112	20.1	26.0	24.0	29.9
29	Difficulty understanding jokes	2.230	1.032	10.8	11.8	21.1	56.4
30	Difficulty understanding slang	2.451	0.964	8.3	8.8	12.3	70.6
31	Difficulty identifying teasing	2.407	0.929	8.3	5.9	22.5	63.2
32	Difficulty understanding when being ridiculed	2.387	0.994	10.3	6.4	17.6	65.7
33	Difficulty understanding why people dislike them	2.480	0.939	8.8	4.9	15.7	70.6
34	Fails to predict social consequences	2.544	0.900	8.3	2.9	14.7	74.0
35	Doesn't seem to understand people have different thoughts and feelings	2.412	0.956	9.3	5.4	20.1	65.2
36	Doesn't understand that the other person doesn't know	2.412	0.976	9.8	5.9	17.6	66.7
37	Needs an excessive amount of reassurance if things are changed	1.637	1.058	16.2	31.9	24.0	27.9
38	Frustrated when they cannot do something	1.976	0.965	9.8	18.1	36.8	35.3
39	Tantrums when frustrated	1.936	1.060	13.7	18.1	28.9	39.2
40	Upset when routines are changed	1.534	1.094	21.6	28.9	24.0	25.5
41	Responds negatively when given commands	1.623	0.941	14.7	26.0	41.7	17.6
42	Has extreme reactions in response to loud, unexpected noise	1.123	1.127	39.7	26.5	15.7	18.1
43	Tantrums when doesn't get their way	1.799	1.075	16.7	19.6	30.9	32.8
44	Tantrums when told to stop something they enjoy	1.691	1.082	19.6	19.6	32.8	27.9
45	Exceptionally precise speech	0.750	0.948	54.9	20.6	19.1	5.4
46	Concrete meanings to words	1.147	1.161	42.2	19.6	19.6	18.6
47	Talks about same thing excessively	1.000	1.132	49.5	15.2	21.1	14.2

Table 9 (cont'd)

48	Superior knowledge in specific subjects	0.569	0.899	65.2	18.6	10.3	5.9
49	Excellent memory	1.054	1.018	38.7	27.5	23.5	10.3
50	Intense, obsessive interest in specific subjects	0.583	0.892	64.2	18.1	12.7	4.9
51	Makes naïve remarks	0.583	0.935	67.6	11.8	15.2	5.4
52	Repeats words or phrases	1.397	1.209	34.8	16.7	22.5	26.0
53	Repeats words out of context	1.299	1.197	38.2	16.2	23.0	22.5
54	Speaks with flat tone, affect	0.809	0.986	52.9	20.1	20.1	6.9
55	Uses "yes" and "no" inappropriately	0.995	1.085	47.5	16.7	24.5	11.3
56	Uses "he" or "she" when referring to self	0.294	0.660	79.9	12.7	5.4	2.0
57	Abnormal speech (tone, volume, rate)	0.976	1.129	51.5	12.7	22.5	13.2
58	Utters idiosyncratic words or phrases	0.976	1.164	52.9	12.3	19.1	15.7

Additionally, Bartlett's Test of Sphericity (Bartlett, 1950) and Kaiser-Meyer-Olkin Test (KMO; Kaiser, 1970) were examined to assess the suitability of the correlation matrix for factor analysis. The Bartlett's test was significant, as desired for suitability, indicating that the correlation matrix is not an identity matrix (χ^2 = 11182.813, *df* = 1653, *p* < 0.001; Raykov & Marcoulides, 2008). The KMO value, at minimum, should be above 0.50 to conduct an EFA. In the study one dataset, the KMO = 0.923. Based on Kaiser and Rice's (1974) KMO standards ranging from < 0.50 (i.e., "unacceptable") to .90 (i.e., "marvelous") including other descriptors in between such as "miserable" for values in the 0.50s, "mediocre" the 0.60s, "middling" for the .70s, and "meritorious" for 0.80s. In the current dataset, the KMO (i.e., 0.923) met the highest standard (i.e., "marvelous"). In all, both the Bartlett's test and KMO indicated suitability for EFA.

GARS-3 communality estimates were also examined for the sample and ranged from 0.301 to 0.912 with an average communality estimate of 0.681. According to the MacCallum et al. article (1999), this would indicate that the dataset had wide communality (i.e., estimates ranging from .2 to .8). Table 1 (p. 93) from the MacCallum et al.'s Monte Carlo simulation was used to estimate the likelihood of a convergent factor structure based on sample size, ratio of items to factors, and communality estimates. Assuming a GARS-3 item-to-factor ratio of approximately 58:6, the closest corresponding table entry was a 20:3 ratio which yielded 100% convergent admissible solutions at a sample size of 60 or above in the simulations (MacCallum, 1999). This strongly supported the adequacy of the sample of 204 participants for the EFA.

Research Question 1

When students with ASD are rated by special education teaching staff with the GARS-3, how many potentially interpretable factors are present and should be considered for retention?

This research question was answered using principal axis factoring, a factoring procedure robust to violations of normality assumptions. Additionally, several indices were used as part of the determination of factor selection. These included the Kaiser criterion, scree plot, parallel analysis, and Velicer's MAP, which were calculated using a combination of SPSS and R statistical packages.

According to the Kaiser criterion (Kaiser, 1960), factors with eigenvalues greater than one should be retained. (This standard originated from principal components analysis [PCA] context, where an eigenvalue of one would be equal to the variance of a single item.) Using the SAS values and eigenvalue greater than 1.0 suggested potentially nine factors. The SPSS values using this standard suggested eight factors, or nine factors if the mean item communality standard (in this case 0.828) was used. (The mean item communality is a logical equivalent of the eigenvalue greater than one standard for non-PCA EFA methods. However, this standard is inconsistently applied in the EFA literature.) Given the discrepancy, the interpretation of nine possible factors was used for purposes of inclusiveness and given the exploratory nature of the analysis. Table 10 shows the eigenvalues for this EFA from both SPSS and SAS outputs.

Table 10. Eigenvalues for the Kaiser Criterion

Factor	Eigenvalues (SPSS R Plugin)	Eigenvalues (SAS)
1	25.233	25.426
2	6.734	6.878
3	4.117	4.250
4	2.753	2.950
5	2.340	2.476
6	1.736	1.983
7	1.180	1.343
8	**1.034**	1.189
9	0.858[a]	**1.035**
10	0.792	0.963
11	0.666	0.824
12	0.590	0.731
13	0.532	0.656
14	0.475	0.620
15	0.419	0.532
16	0.362	0.521
17	0.344	0.489
18	0.311	0.448
19	0.304	0.431
20	0.258	0.382
21	0.240	0.379
22	0.223	0.357
23	0.193	0.351
24	0.169	0.330
25	0.157	0.303
26	0.109	0.227
27	0.090	0.212
28	0.070	0.199
29	0.049	0.193
30	0.043	0.180
31	0.032	0.155
32	0.018	0.151
33	0.003	0.136
34	-0.003	0.124

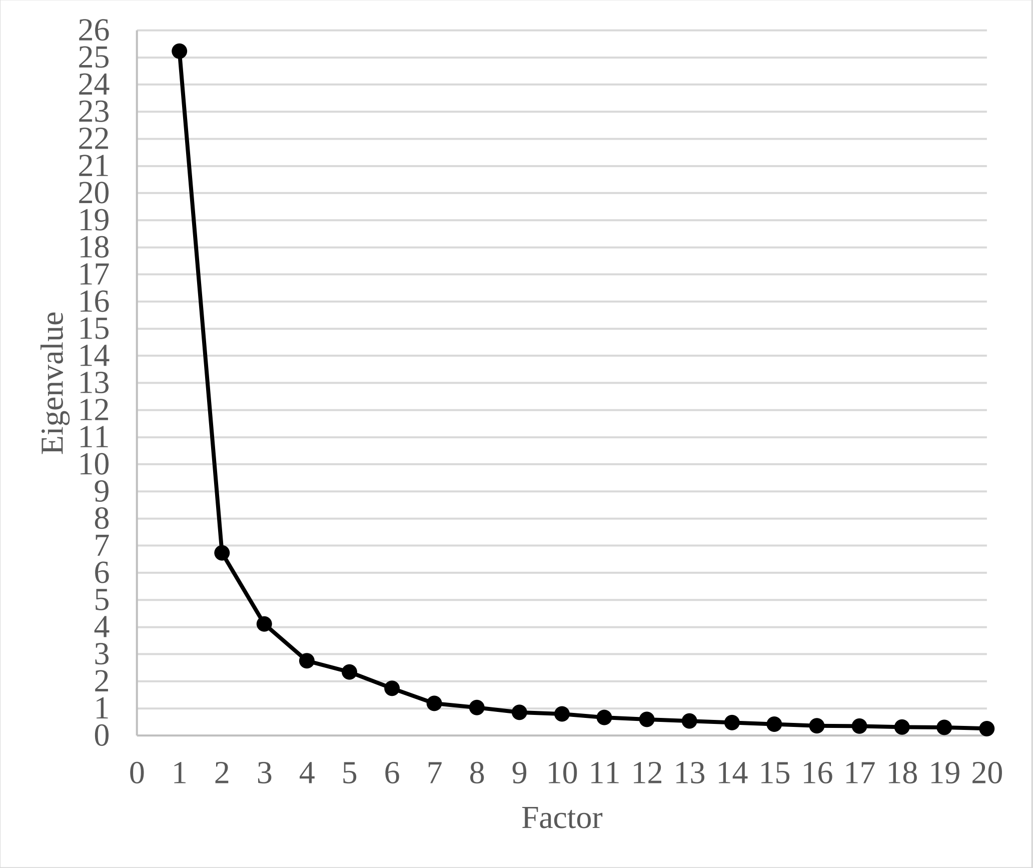
Table 10 (cont'd)

35	-0.010	0.112
36	-0.023	0.091
37	-0.038	0.082
38	-0.041	0.071
39	-0.053	0.054
40	-0.064	0.052
41	-0.074	0.047
42	-0.077	0.033
43	-0.080	0.018
44	-0.084	0.013
45	-0.105	0.005
46	-0.111	0.002
47	-0.113	0.001
48	-0.135	0.001
49	-0.151	0.001
50	-0.158	0.000
51	-0.172	0.000
52	-0.205	0.000
53	-0.281	0.000
54	-0.371	-0.001
55	-0.423	-0.001
56	-0.475	-0.001
57	-0.613	-0.002
58	-0.748	-0.002

[a]The SPSS Ordinal Factor Analysis Menu access through the R Plugin calculates the eigenvalues slightly different from SAS. Importantly, whether or not the Kaiser criterion suggests 8 or 9 factors, is dependent on whether one uses the PCA standard of 1.0 or a common factors equivalent of the mean communality estimate, which in this case was 0.828.

The scree plot test (Cattell, 1966) was also used. Figure 1 shows the scree plot using eigenvalues calculated using the SPSS R Plugin Ordinal Factor Analysis Menu. For purposes of interpretation and succinctness, the first 20 eigenvalues were provided. After determining the "break" in the graph, researchers look at factor solutions with the number of factors above this "break." In the current study, there appears to be two "points of inflection" for eigenvalues at four and seven, indicative of examining the structures with the number of data points prior to these breaks – three- and six- factor structures (Yong & Pearce, 2013).

Figure 1. Scree Plot of Eigenvalues Generated by the SPSS R Plugin Ordinal Factor Analysis Menu



Velicer's MAP (Velicer, 1976) was generated by SPSS R Plugin Ordinal Factor Analysis Menu (Basto & Pereira, 2012). This test examines the systematic and unsystematic variance with different number of factors (O'Connor, 2000; Velicer, 1976). Partial correlations are computed, averaged, and squared with factor retention occurring until the unsystematic variance is greater than the systematic (O'Connor, 2000). Table 11 shows the results from the Velicer's MAP test. The results using the squared average partial correlations indicates retention of seven factors, as the seventh factor has the lowest squared average partial correlation of 0.0287. The results using the fourth average partial correlations, offering a perspective based on another standard, indicate retention of eight factors, as the eighth factor had the lowest fourth average partial correlation of 0.00286.

Table 11. Velicer's MAP Test Values for the Squared Average and 4th Average Partial Correlations

Factors	Squared Average Partial Correlations	4th Average Partial Correlations
0	0.2063	0.08012
1	0.0654	0.01615
2	0.0538	0.01139
3	0.0481	0.00892
4	0.0425	0.00666
5	0.0391	0.00548
6	0.0290	0.00302
7	**0.0287**	0.00288
8	0.0289	**0.00286**
9	0.0292	0.00298
10	0.0298	0.00298
11	0.0310	0.00318
12	0.0317	0.00339
13	0.0327	0.00353
14	0.0340	0.00379
15	0.0358	0.00413
16	0.0377	0.00425
17	0.0391	0.00467
18	0.0414	0.00510
19	0.0432	0.00580
20	0.0454	0.00618
21	0.0475	0.00666
22	0.0501	0.00734
23	0.0525	0.00799
24	0.0533	0.00777
25	0.0537	0.00779
26	0.0565	0.00863
27	0.0616	0.01046
28	0.0669	0.01260
29	0.0732	0.01418
30	0.0784	0.01658
31	0.0842	0.01830
32	0.0910	0.02130
33	0.0965	0.02357
34	0.1045	0.02757
35	0.1172	0.03433
36	0.1327	0.04263
37	0.1502	0.05342

Table 11 (cont'd)

38	0.1716	0.06620
39	0.2027	0.08807
40	0.2563	0.13238
41	0.3073	0.17923
42	0.3608	0.22373
43	0.5203	0.39225
44	0.9441	0.92149
45	0.0746	0.01575
46	0.0841	0.01919
47	0.0942	0.02335
48	0.1063	0.02956
49	0.1209	0.03630
50	0.1389	0.04594
51	0.1520	0.05271
52	0.1765	0.06840
53	0.2069	0.09059
54	0.2549	0.13035
55	0.3260	0.19500
56	0.4965	0.37035

Parallel analysis (Horn, 1965) was also calculated through the R Plugin via SPSS. Parallel analysis seeks to compare the variance of the model to random variance (Horn, 1965; O'Connor, 2000). Random data is used to compute eigenvalues and matrices, with factor retention occurring until the random data has a greater eigenvalue than the target data set for the same number of factors (O'Connor, 2000). Table 12 shows the results of the parallel analysis and indicates retention of five factors.

Table 12. Parallel Analysis Values for Obtained and Random Variance Eigenvalues

Factor	Observed	Parallel Analysis
	Eigenvalues	Eigenvalues[a]
1	25.233	2.428
2	6.734	2.170
3	4.117	2.055
4	2.753	1.989
5	**2.340**	**1.882**
6	1.736	1.836
7	1.180	1.733
8	1.034	1.644
9	0.858	1.575
10	0.792	1.499
11	0.666	1.468
12	0.590	1.403
13	0.532	1.305
14	0.475	1.266
15	0.419	1.193
16	0.362	1.139
17	0.344	1.083
18	0.311	1.036
19	0.304	0.976
20	0.258	0.956
21	0.240	0.901
22	0.223	0.879
23	0.193	0.819
24	0.169	0.785
25	0.157	0.757
26	0.109	0.728
27	0.090	0.672
28	0.070	0.619
29	0.049	0.597
30	0.043	0.567
31	0.032	0.549
32	0.018	0.520
33	0.003	0.463
34	-0.003	0.403
35	-0.010	0.390
36	-0.023	0.352
37	-0.038	0.312
38	-0.041	0.273
39	-0.053	0.237
40	-0.064	0.210
41	-0.074	0.196
42	-0.077	0.173

Table 12 (cont'd)

43	-0.080	0.109
44	-0.084	0.095
45	-0.105	0.083
46	-0.111	0.045
47	-0.113	0.003
48	-0.135	-0.026
49	-0.151	-0.033
50	-0.158	-0.079
51	-0.172	-0.093
52	-0.205	-0.105
53	-0.281	-0.154
54	-0.371	-0.177
55	-0.423	-0.186
56	-0.475	-0.216
57	-0.613	-0.233
58	-0.748	-0.248

[a]The parallel analysis eigenvalues reflect the 95th percentile of the eigenvalue distribution.

The purpose of research question one was to examine the number of potentially interpretable factors using criterion tests of the Kaiser criterion, the scree plot, Velicer's MAP, and parallel analysis. A summary of the results above is found below in Table 13. Broadly, the criterion tests results suggest examining factor solutions consisting of three to nine factors; however, more conservative criteria (i.e., Velicer's MAP and parallel analysis) suggest a range of interpretable factors between five and eight factors.

Table 13. Summary of Factor Retention Criterion Tests

Criterion	Suggested Number of Factors to Retain
Kaiser Criterion	8, 9
Scree Test	3, 6
Velicer's MAP Test	7, (8)[a]
Parallel Analysis	5

[a]For the Velicer's MAP test, the squared value was primarily relied upon, but the value associated with the fourth average squared partial correlation was also reported for comprehensiveness.

Research Question 2

When students with ASD are rated by special education teaching staff with the GARS-3, how many factors should be retained to yield the most interpretable factor solution for the GARS-3?

In order to determine the number of interpretable factors, the pattern matrices were examined for the solutions with the number of factors suggested by the previously stated factor retention criterion tests (see Table 13). It was most reasonable to rely primarily on Velicer's MAP and parallel analysis criteria, as these are considered the more reliable and accurate factor selection criteria. However, given the exploratory nature of this analysis, the criteria were expanded to include other factor selection criteria to generate a range of factor solution options. Velicer's MAP and parallel analysis suggested five, seven, or eight factors and prioritized looking at solutions with one more. (In the interest of taking sampling error into account, to at least some extent, solutions consisting of one fewer and one more factor than these specific tests suggested were also assessed for interpretability. This is why a range of solutions are listed [i.e., solutions with between five and eight factors]).). As indicated above, given the exploratory nature of the analysis and use of additional factor selection criteria (e.g., Kaiser criterion, scree plot), a broader range of solutions were also examined (i.e., solutions with between two and ten factors). A promax, oblique rotation, was used to account for correlated constructs within the data (Costello & Osborne, 2005).

Interpretation. The most conservative criterion tests for factor retention suggested interpretable factors ranging between five and eight. Within the five-factor solution, factor one broadly encompassed a social construct, including aspects such as social understanding, interaction, and communication. Items that loaded most highly on factor two related to stereotyped or repetitive behaviors and interests. Items that loaded highest on factor three encompassed constructs related to emotional and behavioral inflexibility. The fourth factor was specific to rigid thought patterns and restricted and repetitive interests. The fifth factor included items related to stereotyped or repetitive speech and communication.

The six-factor solution included four of the five factors observed in the five-factor solution (i.e., related to stereotyped or repetitive behaviors and interests, emotional and behavioral inflexibility, rigid thought patterns, and stereotyped or repetitive speech/communication). Compared to the five-factor solution, the factor related to social constructs split into two factors in the six-factor solution. The first factor in the six-factor solution was specific to social interactions (e.g., interest, initiation, reciprocity). The other social factor, now factor four, was specific to understanding within a social context.

The seven-factor solution included five of the six factors observed in the prior six-factor solution. The factor related to restrictive and/or repetitive behavior and interests in the six-factor solution split into two factors in the seven-factor solution. One factor that emerged encompassed restricted and repetitive behaviors (RRBs), but seemed to focus on stereotyped movements (e.g., hand flapping, finger flicking, rapid lunging). The other factor that emerged related to restrictive and repetitive behavior and interests, but items appeared to be situation-specific to free time or play (e.g., using toys inappropriately, stereotyped or repetitive behavior during play) in addition to sensory aspects of behavior (e.g., during play, in relation to body parts, etc.).

The eight-factor solution was deemed un-interpretable as the eighth factor consisted of only one item, and the other seven factors were the same as those observed in the seven-factor model. The item made up the eighth factor came from the factor – within the seven-factor solution – related to stereotyped or repetitive speech and communication. Factor solutions above eight factors (i.e., the nine- and ten-factor solutions) were subject to the same issue (e.g., only one item uniquely loaded on a factor, no items loaded uniquely on a factor, etc.).

For thoroughness, two-, three-, and four-factor solutions within the range of other factor retention indicators were also examined. The two-factor solution consisted of one factor largely pertaining to social-related items and another factor representing the remaining constructs (e.g., RRBs, emotion, language, cognition, etc.). The three-factor solution yielded a large factor related to social interaction and communication in addition to RRBs, a factor related to emotional responding, and a third factor pertaining to language, cognition, and rigidity. These factor solutions were difficult to interpret as the obtained factors often appeared to aggregate across multiple constructs that were difficult to capture conceptually as a general construct or were simply too general and could conceivably be refined into separate and meaningful factors. In the four-factor solution, three very broad factors emerged (i.e., a broad social construct, stereotyped or repetitive behaviors and interests, emotional and behavioral inflexibility). Factor four appeared difficult to interpret as it measured multiple constructs– it included items that related to stereotyped speech and communication, restricted interests, and rigidity in thought patterns. Of the twelve items on factor four, four items yielded substantive item loadings (i.e., loadings >.30).

Four additional researchers, with a background in ASD and measurement, independently examined the wide range of solutions. Each researcher concluded that the six-factor solution was most interpretable. The researchers discussed solutions and specifically highlighted solutions with five to seven factors. These solutions seemed most interpretable, and the solutions captured similar constructs. The five-factor solution contained one large social construct that was conceptually relevant but seemed to measure multiple constructs. This large social factor also yielded substantive cross-loadings for two of the items. This five-factor solution, overall, had six items across its factors with substantive cross-loadings.

In the six-factor solution, the larger social construct from the five-factor solution was divided into two specific social constructs– one factor related to social interest and reciprocity while the other factor captured social nuances and understanding. There were two substantive cross-loadings on one of these factors and one on the other. All of these cross-loadings related to the other social factor. This contrasts with the five-factor solution where the social factor had substantive cross-loadings with a factor that did not predominantly relate to a social construct (i.e., stereotyped or repetitive speech and communication). Overall, this solution had eight substantive cross-loadings among eight different items across the factors.

The seven-factor solution retained the two distinct social factors from the six-factor solution. However, the factor related to RRBs was split: one factor specific to these behaviors during free play and another factor regarding specific stereotyped movements. The former factor, related to free play, contained seven items – five which had substantive cross-loadings with other factors. The latter factor related to stereotyped movements contained six items – two of which had substantive cross-loadings with other factors. Across all seven factors, there were thirteen items with substantive cross-loadings.

As mentioned, the researchers concluded the six-factor model was most interpretable. They discussed construct relevance of the factors in addition to the patterns of substantive cross-loadings. The six-factor solution contained factors with clear, interpretable constructs with fewer cross-loadings (e.g., as compared to the seven-factor solution).

The Retained, Most Interpretable Solution: Six-Factor Model. After concluding that the six-factor model was most interpretable, these same researchers independently named the factors and then discussed their interpretations and names for each factor in each solution. For many factors, there were near identical or overlapping names (e.g., social interest/reciprocity, restricted and repetitive/stereotyped behaviors, emotion regulation, social understanding/nuances, atypical/stereotyped speech/language) across the researchers. However, one factor in particular was more difficult to name as it was more complex and multifaceted. Specifically, the researchers had difficulty identifying a succinct name that encompassed or captured the entirety of the factor. This fifth factor had items related to restricted interests, rigidity or atypical thought patterns, and processing/cognition. In the published factor solution, this scale was named Cognitive Style (Gilliam, 2013), but there was discussion among the researchers about the items representing something more inherent or innate (versus style), which seemed to involve more of a choice.

After in-depth discussion the researchers came to consensus on the assignment of the following factor names: Social-Emotional Reciprocity, Restricted & Repetitive Behaviors, Emotion Regulation, Social Understanding, Cognitive Disposition, Speech & Language. The names or constructs chosen by the researchers were quite similar to the subscale names in the published six-factor solution (i.e., Restricted/Repetitive Behaviors, Social Interaction, Social Communication, Emotional Responses, Cognitive Style, Maladaptive Speech [Gilliam, 2013]). Table 14, below, depicts the retained six-factor solution of study one, with items sorted by factor and descending from the highest to the lowest loading per factor. Please refer to Appendix C for the pattern matrix of the seven-factor solution.

Table 14. Study One Six-Factor Solution Pattern Matrix

Item	Item Stem	Factor					
		1	2	3	4	5	6
19	Shows minimal expressed pleasure in interactions	**.95**	-.08	-.04	.01	.04	.04
23	Shows minimal or no response when others attempt to interact	**.95**	.01	.09	-.14	.01	-.03
27	Shows little or no interest in others	**.94**	.03	.03	-.07	-.01	.01
18	Seems indifferent to other person's attention	**.92**	-.15	-.04	.09	.01	.07
22	Seems unwilling to get others to interact	**.89**	.06	-.03	.04	.05	.01
20	Displays little or no excitement in showing toys or objects	**.86**	-.09	.03	.14	-.01	-.02
21	Seems uninterested in pointing out things	**.80**	.04	-.09	.16	-.08	-.03
25	Doesn't try to make friends	**.80**	.12	-.01	.11	-.06	-.04
15	Pays little or no attention to peers	**.78**	.10	-.06	.07	.00	.09
24	Displays little or no reciprocal communication	**.73**	.05	.06	.06	-.13	-.03
16	Fails to imitate	**.69**	.09	.02	.10	-.03	.01
14	Does not initiate conversations	**.64**	-.07	.01	.36	-.05	-.03
26	Fails to engage in creative play	**.61**	.02	-.04	.29	-.16	-.11
17	Doesn't follow other's gestures to look at something	**.61**	.15	.00	.14	-.02	.01
28	Responds inappropriately to humorous stimuli	**.45**	.07	.01	.39	.02	.01
6	Flap hands or fingers	-.18	**.91**	-.18	.10	-.04	-.08
4	Flicks fingers rapidly in front of eyes	-.13	**.87**	-.11	.08	-.04	-.02
10	Engages in stereotyped behaviors in play	.06	**.82**	.06	.01	.06	.05
3	Stares at hands, objects, or items in environment	.22	**.74**	-.04	-.07	.01	.05
5	Makes rapid lunging, darting movements	-.07	**.74**	.10	.12	-.06	-.14
7	Makes high-pitched sounds or other vocalizations	.19	**.68**	-.01	-.04	-.27	-.04
9	Does things repetitively	.06	**.67**	.17	.22	.26	-.06

Table 14 (cont'd)

1	Majority of time alone spent in repetitive or stereotyped behaviors	.39	**.66**	-.02	-.14	.23	.01
13	Ritualistic or compulsive behaviors	-.05	**.64**	.22	.26	.18	-.05
8	Uses toys or objects inappropriately	.07	**.63**	.12	.14	-.03	-.05
12	Shows unusual interest in sensory aspects	.11	**.63**	.08	.08	-.10	.10
2	Preoccupied with specific stimuli	.42	**.61**	.13	-.29	.08	.12
11	Repeats unintelligible sounds	.09	**.50**	.13	-.08	-.29	.37
43	Tantrums when doesn't get their way	-.17	-.04	**.99**	.04	-.11	.07
39	Tantrums when frustrated	-.05	.09	**.95**	-.07	-.09	.00
41	Responds negatively when given commands	-.03	-.03	**.93**	.00	-.01	-.10
44	Tantrums when told to stop something they enjoy	-.10	.00	**.93**	-.04	-.16	.11
40	Upset when routines are changed	.06	.03	**.83**	.01	.11	-.01
38	Frustrated when they cannot do something	.23	.00	**.82**	-.11	.18	-.07
37	Needs an excessive amount of reassurance if things are changed	.06	-.10	**.80**	.23	.23	-.09
42	Has extreme reactions in response to loud, unexpected noise	.13	.05	**.62**	-.01	-.14	.05
33	Difficulty understanding why people dislike them	.16	.13	-.07	**.84**	.07	-.05
34	Fails to predict social consequences	.21	-.04	.06	**.82**	.06	.05
30	Difficulty understanding slang	.23	.01	.07	**.78**	.01	.00
31	Difficulty identifying teasing	.23	.08	-.06	**.77**	.01	.05
32	Difficulty understanding when being ridiculed	.24	.14	-.14	**.76**	.01	.01
36	Doesn't understand that the other person doesn't know	.16	.00	.08	**.76**	.00	.12
35	Doesn't seem to understand people have different thoughts and feelings	.18	.02	.14	**.69**	-.08	.15
29	Difficulty understanding jokes	.32	-.03	.09	**.67**	-.10	.01
50	Intense, obsessive interest in specific subjects	.16	.00	.02	-.12	**.92**	-.09
48	Superior knowledge in specific subjects	.04	.04	-.09	-.07	**.91**	-.04

Table 14 (cont'd)

46	Concrete meanings to words	-.01	-.10	.12	.18	**.81**	-.06
49	Excellent memory	-.16	.00	.01	.11	**.80**	.01
45	Exceptionally precise speech	-.08	-.05	-.17	-.10	**.74**	.15
51	Makes naïve remarks	-.16	.04	-.02	.02	**.74**	.09
47	Talks about same thing excessively	-.09	.09	.16	-.04	**.62**	.36
53	Repeats words out of context	.05	-.04	-.05	-.04	.11	**.89**
52	Repeats words or phrases	.00	.01	-.14	.11	.05	**.84**
58	Utters idiosyncratic words or phrases	.00	.02	.10	-.01	-.09	**.83**
57	Abnormal speech (tone, volume, rate)	-.03	-.03	.08	.16	.01	**.72**
54	Speaks with flat tone, affect	.23	-.19	.00	.12	.11	**.69**
56	Uses "he" or "she" when referring to self	-.14	.07	.01	-.06	.29	**.49**
55	Uses "yes" and "no" inappropriately	-.05	.12	-.12	.36	-.10	**.43**

Note. Loadings assigned to each individual factor are bolded. Loadings greater than .30 are underlined.

Factor 1: Social-Emotional Reciprocity. The first factor in this solution related to social-emotional reciprocity and consisted of fifteen items (i.e., items 14 through 28). Three of the highest item loadings pertained to minimal expressed pleasure during interactions (item 19, .95), minimal or lack of response when others interact (item 23, .95), and little or no interest in others (item 27, .94). Two items had cross-loadings (i.e., loadings ≥ .30 on another factor) with factor 4 (Social Understanding): item 14, related to a lack of initiation in conversations (.64), had a cross-loading of .36 and item 28, related to inappropriate responses to humorous stimuli (.45), had a cross-loading of .39.

Factor 2: Restricted & Repetitive Behaviors. The second factor, related to restricted and repetitive behaviors, was comprised of thirteen items (i.e., items 1 through 13). The highest loadings on this factor were item 6 (.91) related to hand or finger flapping, item 4 (.87) related to finger flicking in front of eyes, and item 10 (.82) related to stereotyped behaviors with toys or objects during play. Two items cross-loaded with factor 1 (Social-Emotional Reciprocity): item 1 (.66), related to alone time spent in repetitive or stereotyped behaviors, had a cross-loading of .39 and item 2 (.50), related to an intense preoccupation with specific stimuli, had a cross-loading of .42. Item 11 (.50), related to repetitive unintelligible sounds, also had a cross-loading (.37) with factor 6 (Speech & Language).

Factor 3: Emotion Regulation. Factor 3, related to emotion regulation, was comprised of eight items (i.e., items 37 through 44). The items with loadings greater than .82 all pertained to temper tantrums/becoming upset such as when an individual does not get their way (item 43, .99), when frustrated (item 39, .95), when given a directive (item 41, .39), when told to stop doing something enjoyable (item 44, .93), when routines are changed (item 38, .83), and when told they cannot do something (item 37, .82). There are no item cross-loadings ≥ .30 on this factor.

Factor 4: Social Understanding. The fourth factor, related to social understanding, is composed of eight items (i.e., items 29 through 36). The items in this factor all pertained to an aspect of understanding or perspective-taking within a social context such as with item 33 (.84) related to understanding why people may dislike an individual, item 34 (.82) related to the inability to predict consequences in social situations, and item 30 (.78) related to understanding slang expressions. One cross-loading occurred for item 29 (.67), related to understanding jokes, with factor 1 (Social-Emotional Reciprocity; .32).

Factor 5: Cognitive Disposition. Factor 5, which pertained to aspects of cognition, consisted of seven items (i.e., items 45 through 51). The three highest loading items on this factor were item 50 (.92) related to an intense interest in a specific subject, item 48 (.91) related to superior knowledge in a specific subject, and item 46 (.81) related to concrete meanings for words. The only cross-loading that occurred was with item 47 (.62), related to excessively talking about one subject, with factor 6 (Speech & Language; .36).

Factor 6: Speech & Language. The sixth factor, related to atypical speech and language, was comprised of seven items (i.e., items 52 through 58). These items all connected to either the type of speech used (i.e., repetitive, idiosyncratic) or the way language is used (e.g., incorrect pronouns). The three items that loaded highest were related to repeating words out of context (item 53, .89), repeating words or phrases (item 52, .84), and using words or phrases that have no meaning to other people (i.e., idiosyncratic; item 58, .83). One cross-loading occurred with factor 4 (Social Interaction); item 55 (.43), related to using "yes" or "no" incorrectly, had a loading of .36 on factor 4.

In all, research question two sought to determine how many factors should be retained to yield the most interpretable factor solution for the GARS-3 with the current study's sample of students with ASD rated by special education teaching staff. Results from the EFA indicated that a six-factor solution was most interpretable and meaningful.

Research Question 3

When students with ASD are rated by special education teaching staff with the GARS-3, are there substantive correlations between at least some GARS-3 factors within the most interpretable factor structure?

Hypothesis 3. Within the most interpretable factor structure, correlations between at least some factors will be ≥ .30.

After the six-factor solution was retained, correlations between factors were examined to determine whether there were substantive correlations between at least some of the GARS-3 factors. In the table below, the inter-factor correlations of the six-factor solution are presented.

Table 15. Study One Six-Factor Solution Inter-Factor Correlation Matrix

Factor Number and Name	Factor Number					
	1	2	3	4	5	6
Factor 1: Social-Emotional Reciprocity	1.00					
Factor 2: Restricted & Repetitive Behaviors	**.56**	1.00				
Factor 3: Emotion Regulation	**.35**	**.49**	1.00			
Factor 4: Social Understanding	**.59**	**.44**	**.37**	1.00		
Factor 5: Cognitive Disposition	-.29	-.11	-.01	-.26	1.00	
Factor 6: Speech & Language	**.32**	**.40**	**.30**	**.30**	.14	1.00

Note. Correlations greater than or equal to .30 are bolded.

As shown in Table 15, five of the six factors were substantively correlated with each other (i.e., ≥ .30). Using Cohen's (1988) interpretive standards for correlation coefficients, correlations ≥ .50 are considered to be large, and correlations < .50 and ≥ .30 are considered moderate. Factor 1 (Social-Emotional Reciprocity) had a large correlation with factor 2 (Restricted & Repetitive Behaviors; .56) and factor 4 (Social Understanding; .59), as well as moderate correlations with factor 3 (Emotion Regulation; .35) and factor 6 (Speech & Language; .32). Factor 2 (Restricted & Repetitive Behaviors) was moderately correlated with factor 3 (Emotion Regulation; .49), factor 4 (Social Understanding; .44), and factor 6 (Speech &

Language; .40). Factor 3 (Emotion Regulation) also had moderate correlations factor 4 (Social Understanding; .37) and factor 6 (Speech & Language; .30). Additionally, factor 4 (Social Understanding) was moderately correlated with factor 6 (Speech & Language; .30). Factor 5 (Cognitive Disposition) did not substantively correlate with any other factors.

In conclusion, the hypothesis for research question three was supported as five of the six factors had inter-factor correlations greater than .30.

Internal Consistency Reliability

Beyond inter-factor correlations, the internal consistency reliability estimates were examined for each of the six subscales based on the six factors from the EFA. Both Cronbach's alpha and ordinal alpha were estimated. Table 16 depicts these internal consistency reliability estimates.

Table 16. Internal Consistency Reliability Estimates

Factor	Name	Number of Items	Cronbach's Alpha	Ordinal Alpha
1	Social-Emotional Reciprocity	15	.965	.977
2	Restricted & Repetitive Behaviors	13	.933	.954
3	Emotion Regulation	8	.970	.986
4	Social Understanding	8	.934	.955
5	Cognitive Disposition	7	.865	.919
6	Speech & Language	7	.821	.867

As shown above, Cronbach's alpha estimates ranged from .821 to .970 and ordinal alpha estimates ranged from .867 to .986. According to Salvia et al. (2017) all scales met criteria for weekly monitoring (≥.70) and screening (≥ .80). Using Cronbach's alpha, four of six scales met criteria for individual decision making (≥ .90), while using ordinal alpha, five of six scales met this standard (Salvia et al., 2017). Further, all scales met the .70 reliability standard for research (Nunnally, 1978). When examining reliability estimates with criteria from Murphy and

Davidshofer (2005), as cited in Sattler (2008), the majority of scales fell in the excellent range (i.e., .90 - .99) and those that did not, still fell in the moderately high or good range (i.e., .80 - .89).

Research Question 4

When students with ASD are rated by special education teaching staff with the GARS-3, how does the six-factor EFA solution correspond to the six GARS-3 subscales proposed by the author (Gilliam, 2013)?

Although the names of the factors in the present study do not perfectly align with the names of the published GARS-3 subscales, they overlap substantively in terms of the names and constructs represented. See Table 17 for a comparison of the factor names between models. The current factor names were based on input from the independent researchers who evaluated and interpreted the different factor solutions.

Overall, the model from the EFA and the published GARS-3 model were almost identical, with 57 of the 58 items within the EFA loading on the same factors as in the published model. The one item that loaded on a different factor was Item 28. This item (i.e., responds inappropriately to humorous stimuli) loaded on the GARS-3 Social Communication factor (i.e., corresponding to the present study's Social Understanding factor), but in the present study loaded onto what was comparable to the Social Interaction subscale/factor in the published GARS-3 (i.e., corresponding to the present study's Social-Emotional Reciprocity factor). Otherwise, five out of six factors in the present study had 100% item overlap with the GARS-3 published model (i.e., in the GARS-3: Restricted/Repetitive Behaviors, Emotional Responses, Social Communication, Cognitive Style, and Maladaptive Speech). With Item 28 loading on factor 1 (Social-Emotional Reciprocity) in the present study, this factor had 93.33% (14/15) overlapping items with the published corresponding GARS-3 factor/subscale (Social Interaction). Of note, in the present study, is the cross-loading for Item 28, as this item's highest loading was 0.45 (factor 1: Social-Emotional Reciprocity), but it also had a substantive cross-loading of 0.39 on factor 4 (the factor that it loaded on in the published GARS-3 model; Gilliam, 2013). Refer to Table 17 for the comparison between items from each model.

In addition to the present study's model having a near identical structure to the published GARS-3 model, two things are important to note for these results. The first being that the present study's model yielded higher factor loadings for many of the items when compared to the GARS-3 model loadings reported in the manual. See Table 17 below for a comparison between the highest to lowest item loadings between the two factor models. Second, a major difference between the two models was that the present study's model had several items that had cross-loadings (i.e., eight items had secondary loadings ≥ .30), while, according to the GARS-3 manual, the published GARS-3 model did not yield any cross-loadings. See Table 14 for the factor structure of the current study.

In conclusion, the EFA six-factor solution was very similar to the six GARS-3 subscales proposed by the instrument author. Each model yielded six factors, with the only difference in structure being Item 28 loading highest on a different factor than expected based on the published model. Further, the present study yielded higher factor loadings and items with substantive cross-loadings compared to the GARS-3 published model.

Table 17. Highest to Lowest Item Loading for Each Factor Between the Present Study and the Published GARS-3 (Gilliam, 2013)

Present Study				Published GARS-3 (Gilliam, 2013)			
	Highest to Lowest Item Loadings in Present Study				Highest to Lowest Item Loadings in Published GARS-3 (Gilliam, 2013)		
Factor Names	Item	Loading	Item Stem	Factor Names	Item	Loading	Item Stem
Social-Emotional Reciprocity	19	.95	Shows minimal expressed pleasure in interactions	Social Interaction	22	.95	Seems unwilling to get others to interact
	23	.95	Shows minimal or no response when others attempt to interact		23	.92	Shows minimal or no response when others attempt to interact
	27	.94	Shows little or no interest in others		19	.87	Shows minimal expressed pleasure in interactions
	18	.92	Seems indifferent to other person's attention		27	.83	Shows little or no interest in others
	22	.89	Seems unwilling to get others to interact		25	.79	Doesn't try to make friends
	20	.86	Displays little or no excitement in showing toys or objects		18	.71	Seems indifferent to other person's attention
	21	.80	Seems uninterested in pointing out things		20	.71	Displays little or no excitement in showing toys or objects
	25	.80	Doesn't try to make friends		24	.70	Displays little or no reciprocal communication
	15	.78	Pays little or no attention to peers		21	.69	Seems uninterested in pointing out things
	24	.73	Displays little or no reciprocal communication		15	.62	Pays little or no attention to peers
	16	.69	Fails to imitate		14	.59	Does not initiate conversations

Table 17 (cont'd)

	14	.64	Does not initiate conversations		16	.51	Fails to imitate
	26	.61	Fails to engage in creative play		17	.48	Doesn't follow other's gestures to look at something
	17	.61	Doesn't follow other's gestures to look at something		26	.41	Fails to engage in creative play
	28[a]	.45	Responds inappropriately to humorous stimuli				
							93.33% item overlap (14 of 15 items)
Restricted & Repetitive Behaviors	6	.91	Flap hands or fingers	Restricted/ Repetitive Behaviors	1	.80	Majority of time alone spent in repetitive or stereotyped behaviors
	4	.87	Flicks fingers rapidly in front of eyes		2	.80	Preoccupied with specific stimuli
	10	.82	Engages in stereotyped behaviors in play		10	.70	Engages in stereotyped behaviors in play
	3	.74	Stares at hands, objects, or items in environment		9	.69	Does things repetitively
	5	.74	Makes rapid lunging, darting movements		3	.65	Stares at hands, objects, or items in environment
	7	.68	Makes high-pitched sounds or other vocalizations		8	.64	Uses toys or objects inappropriately
	9	.67	Does things repetitively		6	.63	Flap hands or fingers
	1	.66	Majority of time alone spent in repetitive or stereotyped behaviors		4	.60	Flicks fingers rapidly in front of eyes
	13	.64	Ritualistic or compulsive behaviors		13	.60	Ritualistic or compulsive behaviors
	8	.63	Uses toys or objects inappropriately		7	.57	Makes high-pitched sounds or other vocalizations

Table 17 (cont'd)

	12	.63	Shows unusual interest in sensory aspects		12	.56	Shows unusual interest in sensory aspects
	2	.61	Preoccupied with specific stimuli		5	.47	Makes rapid lunging, darting movements
	11	.50	Repeats unintelligible sounds		11	.37	Repeats unintelligible sounds
							100% item overlap (13 of 13 items)
Emotion Regulation	43	.99	Tantrums when doesn't get their way	Emotional Responses	43	.95	Tantrums when doesn't get their way
	39	.95	Tantrums when frustrated		39	.89	Tantrums when frustrated
	41	.93	Responds negatively when given commands		44	.88	Tantrums when told to stop something they enjoy
	44	.93	Tantrums when told to stop something they enjoy		41	.76	Responds negatively when given commands
	40	.83	Upset when routines are changed		38	.65	Frustrated when they cannot do something
	38	.82	Frustrated when they cannot do something		40	.63	Upset when routines are changed
	37	.80	Needs an excessive amount of reassurance if things are changed		42	.47	Has extreme reactions in response to loud, unexpected noise
	42	.62	Has extreme reactions in response to loud, unexpected noise		37	.39	Needs an excessive amount of reassurance if things are changed
							100% item overlap (8 of 8 items)

Table 17 (cont'd)

Social Understanding	33	.84	Difficulty understanding why people dislike them	Social Communication	31	.89	Difficulty identifying teasing
	34	.82	Fails to predict social consequences		32	.87	Difficulty understanding when being ridiculed
	30	.78	Difficulty understanding slang		33	.82	Difficulty understanding why people dislike them
	31	.77	Difficulty identifying teasing		30	.78	Difficulty understanding slang
	32	.76	Difficulty understanding when being ridiculed		34	.71	Fails to predict social consequences
	36	.76	Doesn't understand that the other person doesn't know		29	.62	Difficulty understanding jokes
	35	.69	Doesn't seem to understand people have different thoughts and feelings		28[a]	.44	Responds inappropriately to humorous stimuli
	29	.67	Difficulty understanding jokes		35	.42	Doesn't seem to understand people have different thoughts and feelings
					36	.41	Doesn't understand that the other person doesn't know
							100% item overlap (8 of 8 items)
Cognitive Disposition	50	.92	Intense, obsessive interest in specific subjects	Cognitive Style	48	.89	Superior knowledge in specific subjects
	48	.91	Superior knowledge in specific subjects		50	.88	Intense, obsessive interest in specific subjects
	46	.81	Concrete meanings to words		45	.68	Exceptionally precise speech
	49	.80	Excellent memory		47	.68	Talks about same thing excessively
	45	.74	Exceptionally precise speech		49	.64	Excellent memory
	51	.74	Makes naïve remarks		46	.45	Concrete meanings to words

Table 17 (cont’d)

	47	.62	Talks about same thing excessively		51	.44	Makes naïve remarks
							100% item overlap (7 of 7 items)
Speech & Language	53	.89	Repeats words out of context	Maladaptive Speech	53	.84	Repeats words out of context
	52	.84	Repeats words or phrases		52	.80	Repeats words or phrases
	58	.83	Utters idiosyncratic words or phrases		58	.60	Utters idiosyncratic words or phrases
	57	.72	Abnormal speech (tone, volume, rate)		56	.48	Uses “he” or “she” when referring to self
	54	.69	Speaks with flat tone, affect		54	.43	Speaks with flat tone, affect
	56	.49	Uses “he” or “she” when referring to self		55	.43	Uses “yes” and “no” inappropriately
	55	.43	Uses “yes” and “no” inappropriately		57	.40	Abnormal speech (tone, volume, rate)
							100% item overlap (7 of 7 items)

Note. Percentage of overlapping items was calculated by dividing the number of items that occurred on the factor of the published GARS-3 model by the number of items on the factor of the study one model. (Any extra items on a factor in the EFA compared to the published GARS-3 model were not included for in the calculation of the overlap for that particular factor (i.e., item 28 on the Social Communication factor.)

[a]Item 28 was the only item that had a different primary factor loading between the models in the present study and in the published GARS-3 (Gilliam, 2013).

Study Two: CFA

Model Specification and Identification

The two models tested in the CFA were the published GARS-3 factor model and the factor model from study one. Both models involve 58 items and 6 factors. The models tested to answer the following research questions were categorized as over-identified models – where there were more observable pieces of information than estimated model parameters (i.e., $df > 0$; Brown, 2015). Additionally, in both models, all factor loadings and inter-factor correlations were freely estimated, while factor variances were fixed to 1.0 to provide scaling for latent variables.

Model Estimation and Fit

The polychoric correlation matrix, based on the ordinal item data, was input for model estimation using Mplus Version 8.2 (Muthén & Muthén, 1998-2017). The estimator used was Weighted Least Squares Mean Variance (WLSMV) due to ordinal, non-normal item data. Through the WLSMV estimator, results yielded indices of model fit within the study two dataset. These fit indices were the chi-squared test (χ^2), standardized root mean square residual (SRMR), root mean square error of approximation (RMSEA), comparative fit index (CFI), and Tucker-Lewis index (TLI). Further, two supplementary information indices were calculated using the robust maximum likelihood estimator (MLR). These information indices were the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), which are not available through the WLSMV estimator. Finally, to directly compare the fit of two models, the Mplus DIFFTEST (i.e., a corrected chi-square difference test available when using WLSMV) was used.

Research Question 5

When students with ASD and non-ASD developmental disorders are rated by special education teaching staff with the GARS-3, does the interpretive model proposed by the GARS-3 test author produce a reasonable fit to the confirmatory sample covariance matrix?

Hypothesis 5. It is predicted that the GARS-3 author's proposed six-factor solution will yield a reasonable fit to the confirmatory sample inter-item covariance matrix.

To determine if published GARS-3 factor model reasonably fits the confirmatory sample covariance matrix, the Mplus WLSMV estimator was used. Results of the chi-squared test were significant ($\chi^2 = 2323.440$, $df = 1580$, $p < .001$) and the SRMR = 0.081 indicated a reasonably good fit (e.g., Hu & Bentler [1999] suggested a value ≤ .08 [as cited in Brown, 2015]). The RMSEA = 0.049 also indicated a good fit (e.g., Hu & Bentler [1999] suggested ≤ .06; Browne & Cudeck [1993] suggested < .05). The CFI and TLI additionally indicated good fit as the values exceeded the suggested ≥ .95 (Brown, 2015; Hu & Bentler, 1999; CFI = 0.981; TLI = 0.980). Because the WLSMV does not give estimators such as the AIC and BIC, the model was run using the MLR to provide this additional information to assess model fit when compared to any competing models. Results indicated AIC = 26367.365 and BIC = 26990.747. See Appendix E for a table of parameter estimates and measurement path model, in addition to the inter-factor correlation matrix.

Overall, the published GARS-3 model showed evidence of good fit with the study sample data. Thus, hypothesis five was supported.

Research Question 6

When students with ASD and non-ASD developmental disorders are rated by special education teaching staff with the GARS-3, does the retained factor solution from the study one EFA produce a reasonable fit to the confirmatory sample inter-item covariance matrix?

Hypothesis 6. It is predicted that the retained factor solution from study one will reasonably fit the inter-item covariance matrix from the confirmatory sample.

Using the Mplus WLSMV estimator, the chi-squared test was significant ($\chi^2 = 2361.659$, $df = 1580$, $p < .001$). The SRMR and RMSEA were almost identical to the original GARS-3 model and indicated a good fit (SRMR = 0.081; RMSEA = 0.050). The CFI and TLI were also almost identical to those of the first model, and indicated good fit (CFI = 0.980; TLI = 0.979). The model was also run using the MLR estimator to provide additional information criteria to compare competing models. MLR yielded an AIC = 26338.297 and BIC = 26961.679. Table 18 provides information on the CFA item parameter results for this six-factor model. Figures 2 through 7 provide a visual representation of measurement model path diagrams for each of the factors corresponding to the study one model.

Of particular note, Item 56 (i.e., uses “he” or “she” when referring to self), has a very low loading (parameter estimate = .264), which is further explored in the discussion section. Please see Appendix D for the inter-factor correlation matrix.

In conclusion, the hypothesis from research question six was supported as the study one six-factor solution demonstrated good fit to the confirmatory sample inter-item covariance matrix.

Table 18. CFA Item Parameter Results for Six-Factor Model Retained in Study One

Factor	Item	Item Stem	Parameter Estimate	Standard Error	*t* Statistic	Two-tailed *p*-value	R^2	Residual Variance
Social-Emotional Reciprocity	14	Does not initiate conversations	0.833	0.031	27.033	< 0.001	0.694	0.306
	15	Pays little or no attention to peers	0.892	0.019	45.782	< 0.001	0.795	0.205
	16	Fails to imitate	0.856	0.024	36.370	< 0.001	0.733	0.267
	17	Doesn't follow other's gestures to look at something	0.841	0.024	34.442	< 0.001	0.706	0.294
	18	Seems indifferent to other person's attention	0.858	0.023	37.749	< 0.001	0.736	0.264
	19	Shows minimal expressed pleasure in interactions	0.895	0.017	52.089	< 0.001	0.801	0.199
	20	Displays little or no excitement in showing toys or objects	0.897	0.019	46.721	< 0.001	0.805	0.195
	21	Seems uninterested in pointing out things	0.952	0.013	73.346	< 0.001	0.906	0.094
	22	Seems unwilling to get others to interact	0.916	0.016	58.676	< 0.001	0.838	0.162
	23	Shows minimal or no response when others attempt to interact	0.899	0.017	53.162	< 0.001	0.809	0.191
	24	Displays little or no reciprocal communication	0.835	0.027	30.435	< 0.001	0.698	0.302
	25	Doesn't try to make friends	0.932	0.017	55.803	< 0.001	0.868	0.132
	26	Fails to engage in creative play	0.840	0.028	29.553	< 0.001	0.706	0.294

Table 18 (cont'd)

	27	Shows little or no interest in others	0.927	0.014	65.584	< 0.001	0.859	0.141
	28	Responds inappropriately to humorous stimuli	0.828	0.034	24.487	< 0.001	0.686	0.314
Restricted & Repetitive Behaviors	1	Majority of time alone spent in repetitive or stereotyped behaviors	0.873	0.030	29.449	< 0.001	0.762	0.238
	2	Preoccupied with specific stimuli	0.931	0.018	51.860	< 0.001	0.867	0.133
	3	Stares at hands, objects, or items in environment	0.835	0.032	26.083	< 0.001	0.697	0.303
	4	Flicks fingers rapidly in front of eyes	0.702	0.052	13.607	< 0.001	0.493	0.507
	5	Makes rapid lunging, darting movements	0.782	0.044	17.842	< 0.001	0.612	0.388
	6	Flap hands or fingers	0.697	0.049	14.251	< 0.001	0.485	0.515
	7	Makes high-pitched sounds or other vocalizations	0.774	0.040	19.591	< 0.001	0.599	0.401
	8	Uses toys or objects inappropriately	0.869	0.029	30.339	< 0.001	0.754	0.246
	9	Does things repetitively	0.840	0.031	27.121	< 0.001	0.705	0.295
	10	Engages in stereotyped behaviors in play	0.875	0.028	31.793	< 0.001	0.766	0.234
	11	Repeats unintelligible sounds	0.797	0.036	22.072	< 0.001	0.635	0.365
	12	Shows unusual interest in sensory aspects	0.869	0.027	32.323	< 0.001	0.756	0.244
	13	Ritualistic or compulsive behaviors	0.890	0.029	31.045	< 0.001	0.792	0.208

Table 18 (cont'd)

Emotion Regulation	37	Needs an excessive amount of reassurance if things are changed	0.752	0.043	17.357	< 0.001	0.565	0.435
	38	Frustrated when they cannot do something	0.873	0.026	34.110	< 0.001	0.762	0.238
	39	Tantrums when frustrated	0.930	0.013	71.714	< 0.001	0.865	0.135
	40	Upset when routines are changed	0.901	0.024	36.948	< 0.001	0.812	0.188
	41	Responds negatively when given commands	0.896	0.019	47.554	< 0.001	0.802	0.198
	42	Has extreme reactions in response to loud, unexpected noise	0.775	0.047	16.500	< 0.001	0.601	0.399
	43	Tantrums when doesn't get their way	0.985	0.007	149.626	< 0.001	0.970	0.030
	44	Tantrums when told to stop something they enjoy	0.954	0.010	99.393	< 0.001	0.911	0.089
Social Understanding	29	Difficulty understanding jokes	0.956	0.013	73.523	< 0.001	0.913	0.087
	30	Difficulty understanding slang	0.947	0.012	76.492	< 0.001	0.897	0.103
	31	Difficulty identifying teasing	0.976	0.007	143.986	< 0.001	0.952	0.048
	32	Difficulty understanding when being ridiculed	0.990	0.004	222.588	< 0.001	0.980	0.020
	33	Difficulty understanding why people dislike them	0.981	0.009	110.369	< 0.001	0.963	0.037
	34	Fails to predict social consequences	0.958	0.012	78.198	< 0.001	0.918	0.082

Table 18 (cont'd)

	35	Doesn't seem to understand people have different thoughts and feelings	0.967	0.010	97.596	< 0.001	0.935	0.065
	36	Doesn't understand that the other person doesn't know	0.969	0.009	105.410	< 0.001	0.940	0.060
Cognitive Disposition	45	Exceptionally precise speech	0.713	0.047	15.259	< 0.001	0.509	0.491
	46	Concrete meanings to words	0.710	0.044	16.302	< 0.001	0.505	0.495
	47	Talks about same thing excessively	0.834	0.035	23.660	< 0.001	0.696	0.304
	48	Superior knowledge in specific subjects	0.864	0.036	24.119	< 0.001	0.747	0.253
	49	Excellent memory	0.691	0.047	14.775	< 0.001	0.477	0.523
	50	Intense, obsessive interest in specific subjects	0.867	0.029	30.058	< 0.001	0.752	0.248
	51	Makes naïve remarks	0.822	0.037	22.255	< 0.001	0.676	0.324
Speech & Language	52	Repeats words or phrases	0.877	0.036	24.064	< 0.001	0.769	0.231
	53	Repeats words out of context	0.813	0.034	23.810	< 0.001	0.661	0.339
	54	Speaks with flat tone, affect	0.711	0.059	11.951	< 0.001	0.506	0.494
	55	Uses "yes" and "no" inappropriately	0.735	0.059	12.462	< 0.001	0.540	0.460
	56	Uses "he" or "she" when referring to self	0.264	0.114	2.316	0.021	0.070	0.930

Table 18 (cont'd)

57	Abnormal speech (tone, volume, rate)	0.811	0.051	15.893	< 0.001	0.658	0.342
58	Utters idiosyncratic words or phrases	0.888	0.040	22.326	< 0.001	0.789	0.211

Figure 2. Path Diagram for Study One Six-Factor Model Social-Emotional Reciprocity Factor

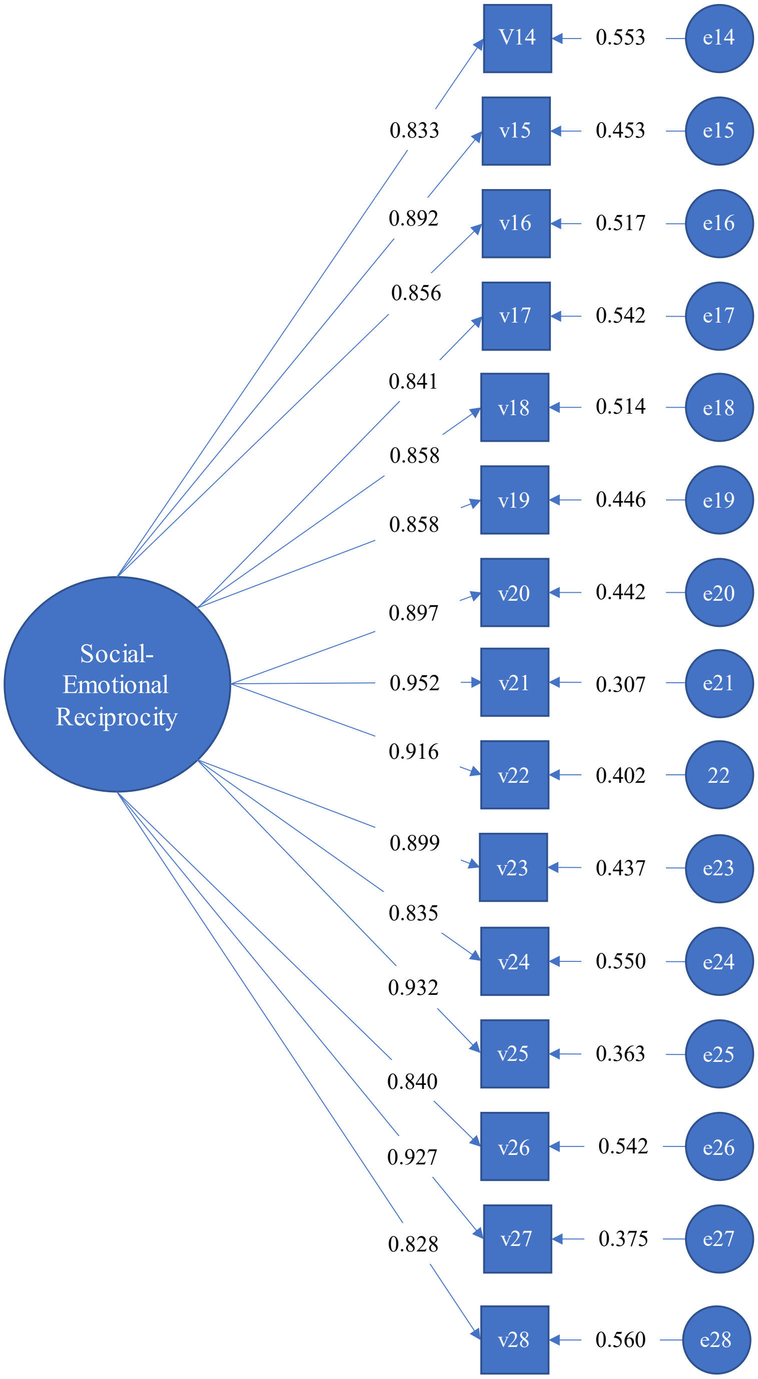


Figure 3. Path Diagram for Study One Six-Factor Model Restricted & Repetitive Behavior Factor

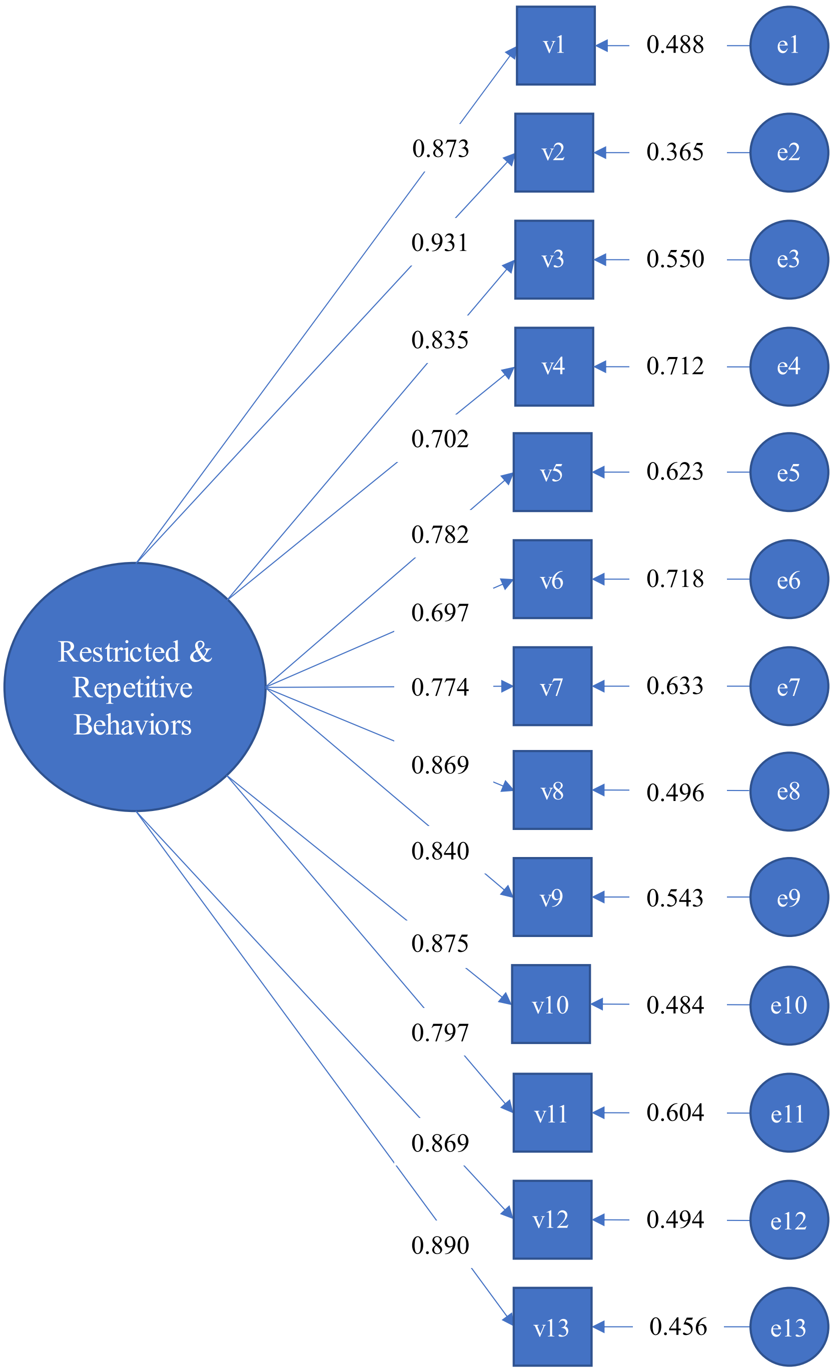


Figure 4. Path Diagram for Study One Six-Factor Model Emotion Regulation Factor

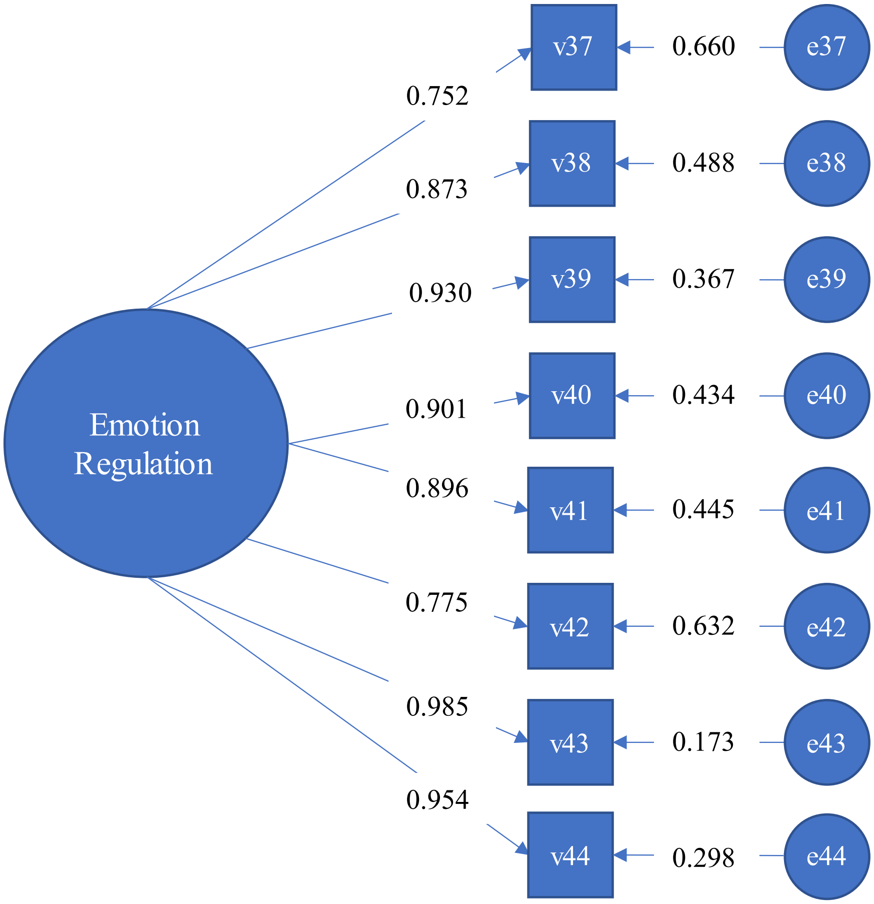


Figure 5. *Path Diagram for Study One Six-Factor Model Social Understanding Factor*

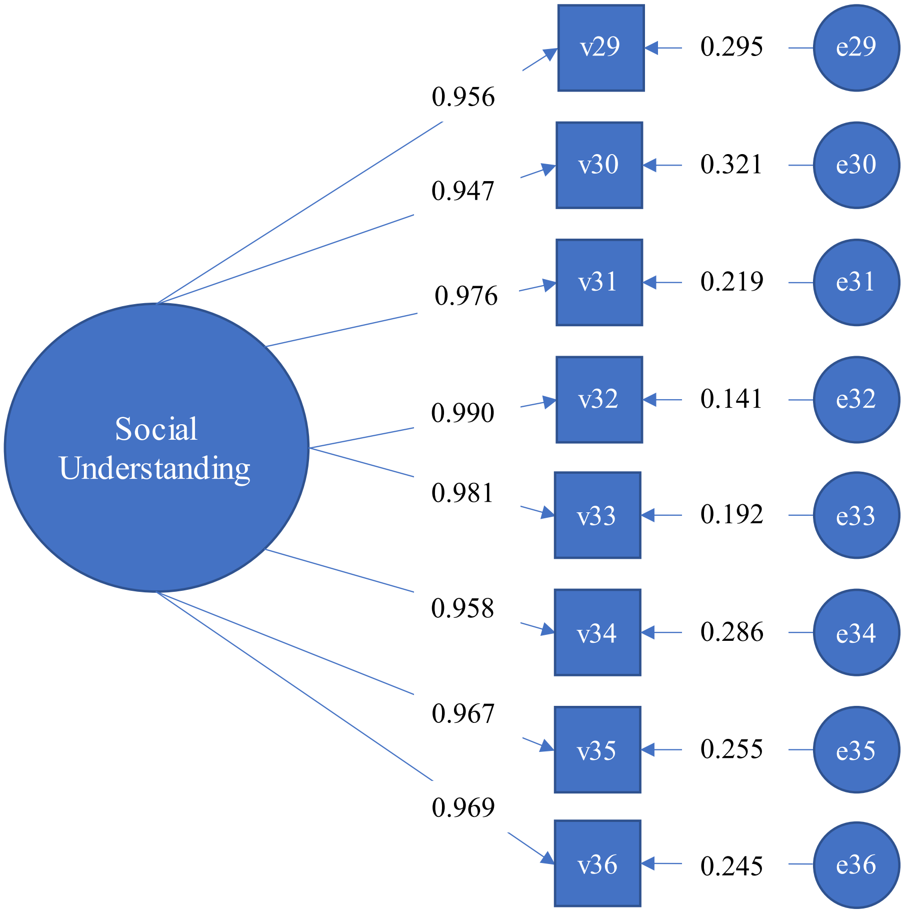


Figure 6. Path Diagram for Study One Six-Factor Model Cognitive Disposition Factor

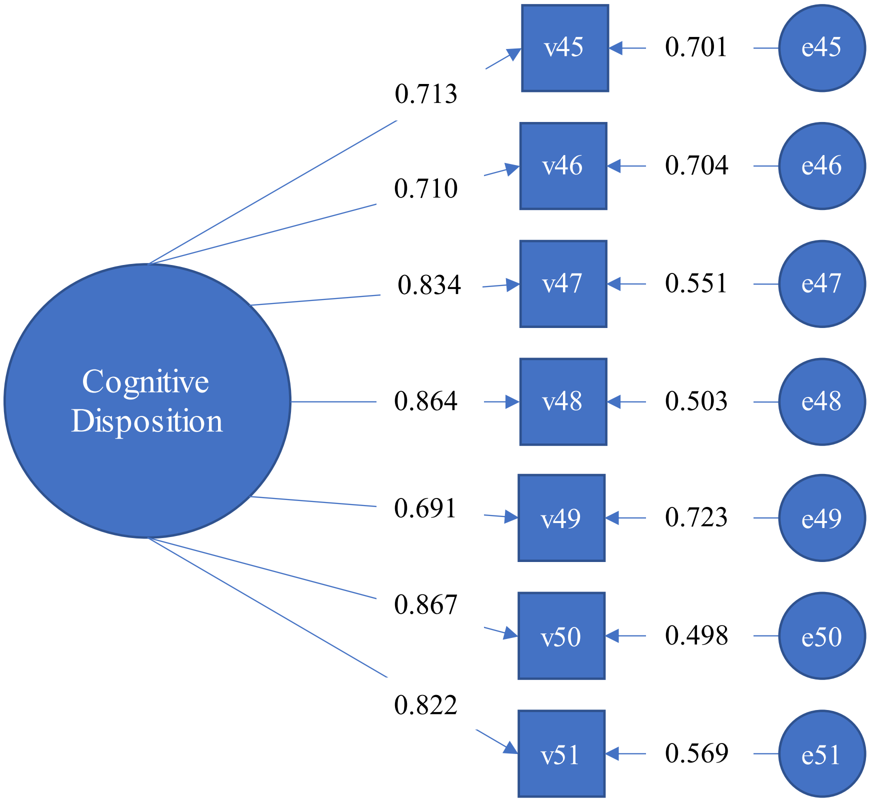
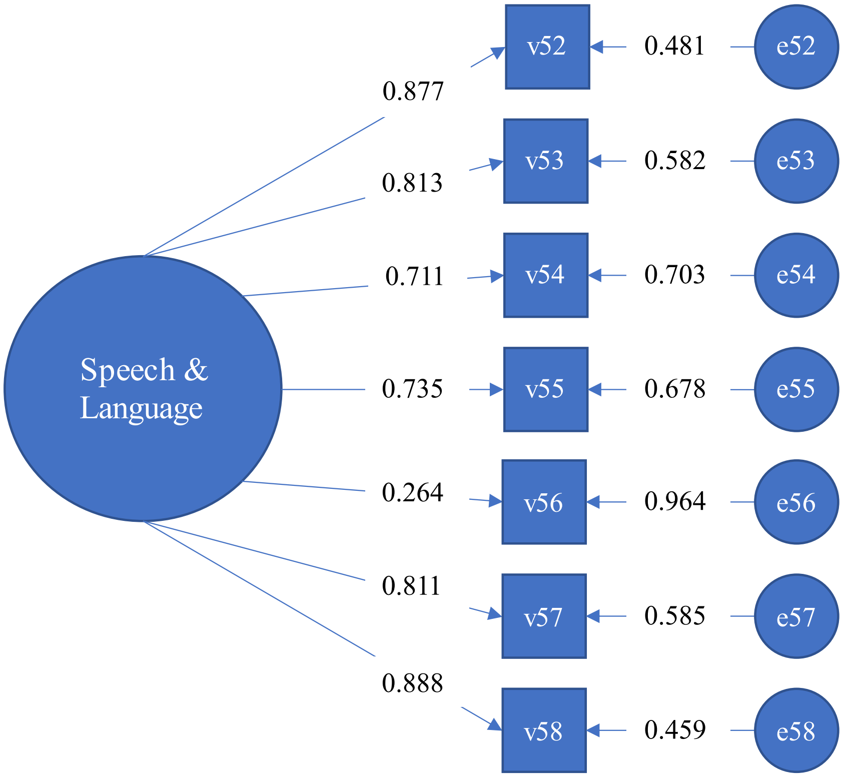


Figure 7. Path Diagram for Study One Six-Factor Model Speech & Language Factor

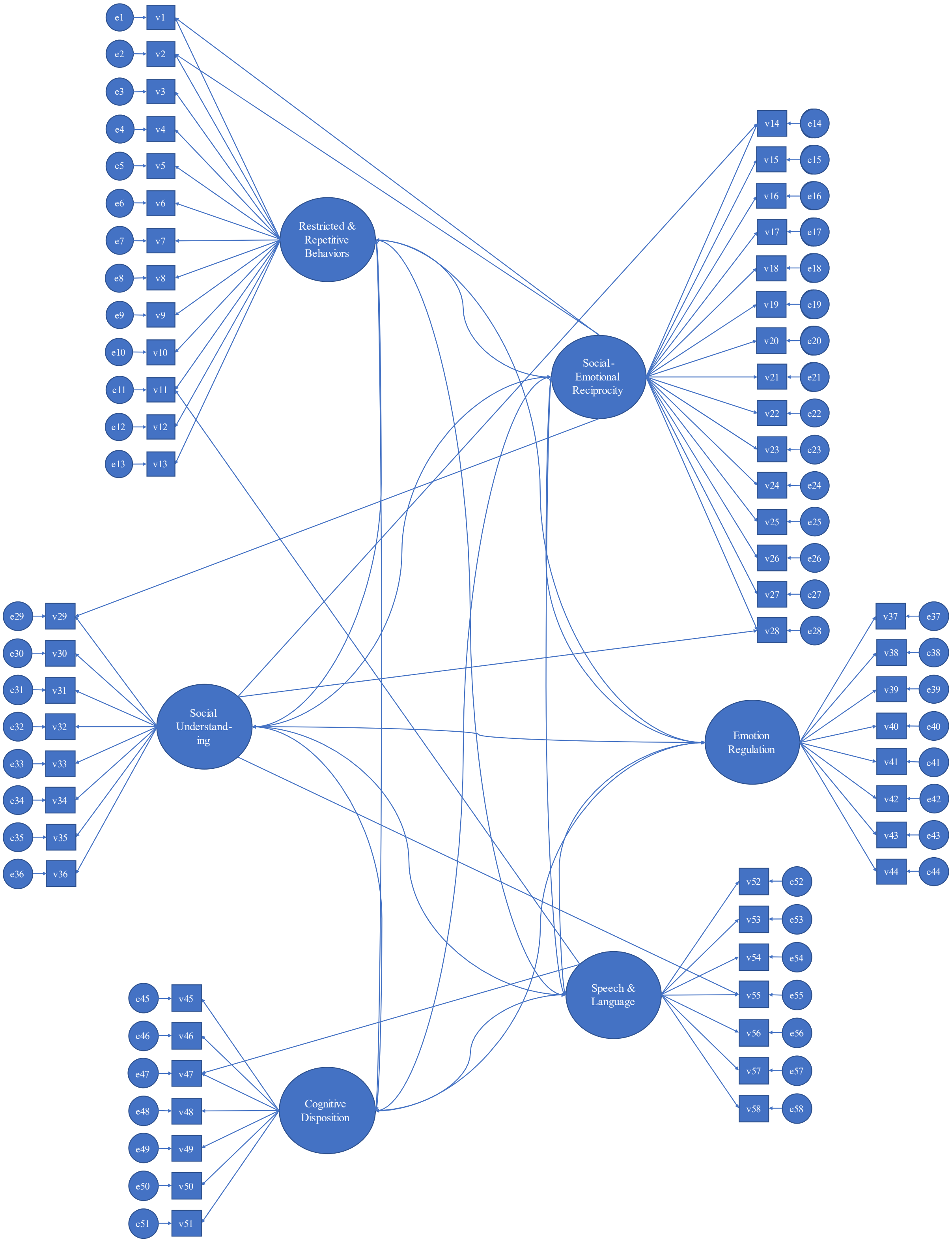


The fit of this model was also tested with the cross-loadings (i.e., ≥ .30) included. While the two models being compared had the same number of factors and almost exactly the same items loading on each factor, the model from study one yielded several cross-loadings, while the instrument manual did not report any for the published GARS-3 factor structure (Gilliam, 2013). Table 19 depicts the factor model with cross-loadings input into Mplus to yield evidence of its fit. When run with the WLSMV indicator, the chi-squared test results were $\chi^2 = 2148.713$, $df = 1572$, $p < .001$. All other indicators met standards of good fit: SRMR = 0.071, RMSEA = 0.043, CFI = 0.985, TLI = 0.984. When tun with the MLR estimator, this model with cross-loadings yielded an AIC = 26242.562 and BIC = 26892.331.

Table 19. Cross-Loadings for Six-Factor Model from Study One

Factor	Items	Cross-Loadings
Social-Emotional Reciprocity	14-28	Item 1: Majority of time alone spent in repetitive or stereotyped behaviors Item 2: Preoccupied with specific stimuli Item 29: Difficulty understanding jokes
Restricted & Repetitive Behavior	1-13	
Emotion Regulation	37-44	
Social Understanding	29-36	Item 14: Does not initiate conversations Item 28: Responds inappropriately to humorous stimuli Item 55: Uses “yes” and “no” inappropriately
Cognitive Disposition	45-51	
Speech & Language	52-58	Item 11: Repeats unintelligible sounds Item 47: Talks about same thing excessively

Figure 8. Path Diagram of Study One Six-Factor Model with Cross-Loadings



Research Question 7

When students with ASD and non-ASD developmental disorders are rated by special education teaching staff with the GARS-3, and the GARS-3 author-proposed model and the EFA-generated model from study one are compared, does one model show evidence of better fit to the confirmatory sample inter-item covariance matrix?

Hypothesis 7. When compared to the author-proposed model, it is predicted that the EFA-generated model from study one will show a substantively better fit with the inter-item covariance matrix from the confirmatory sample.

While research question seven sought to compare these models, the models were extremely similar and had the same degrees of freedom. Therefore, it was not possible to test for a difference in fit using the Mplus DIFFTEST. Given the above results (i.e., under research questions five and six) for RMSEA, SRMR, CFI, TLI, AIC, and BIC, the researcher concluded that both the model from study one and the published GARS-3 model fit similarly. Both are six-factor models with the only difference being the placement of one item. Looking specifically at the AIC and BIC, the model from study one has lower values for these criterion measures, indicative of a slightly better absolute fit (Study One Model: AIC = 26338.297 and BIC = 26961.679; Gilliam Model: AIC = 26367.365 and BIC = 26990.747). Table 20 facilitates the direct comparison of the published GARS-3 model and the study one six-factor model, in addition to the study one model with cross-loadings included.

Table 20. Study Two CFA Model Comparisons by Indicators of Fit

Indicator of Fit	GARS-3 Published Six-Factor Model (Gilliam, 2013)	Study One Six-Factor Model	Study One Six-Factor Model with Cross-loadings
χ^2	2323.440[a]	2361.659[a]	2148.713[a]
SRMR	0.081	0.081	0. 071
RMSEA	0.049	0.050	0.043
CFI	0.981	0.980	0.985
TLI	0.980	0.979	0.984
AIC	26367.365	26338.297	26242.562
BIC	26990.747	26961.679	26892.331

Note. χ^2 = Chi-squared test; SRMR = Standardized root mean square residual; RMSEA = Root mean square error of approximation; CFI = Comparative fit index; TLI = Tucker-Lewis index; AIC = Akaike information criterion; BIC = Bayesian information criterion.

[a] $p < .001$

In conclusion, the hypothesis for research question seven was not supported because the differences between the two models were marginal—with AIC and BIC values slightly favoring the study one model. However, overall, neither model showed evidence of substantially better fit than the other.

Additional Analyses. While conducting an inferential DIFFTEST to statistically compare the fit of the two primary GARS-3 models was not possible given their identical degrees of freedom, additional analyses were conducted to compare the fit of the model from study one (without modeled cross-loadings) to the model from study one with cross loadings included (mentioned above and depicted in Table 19 and Figure 8). These models were nested and use of the DIFFTEST was possible to compare them. Results for the DIFFTEST chi-square were as follows: χ^2 = 112.474, *df* = 8, *p* < .001. The significant results indicated that the study one model with the cross-loadings – with the least number of restrictions, but with the most parameters – fit significantly better, than the model without cross-loadings, within the study two sample data (Kim et al., 2021).

Study Three: Clinical Discriminant Validity

This study explored how well the GARS-3 discriminated between individuals with ASD and individuals with other developmental disabilities requiring substantial support, when rated by special education teaching staff. The examination of clinical discriminant validity involved the assessment of mean differences between the two clinical groups, the sensitivity and specificity of the recommended cut scores for accurately identifying those at risk and those not at risk using the Autism Index 6, and exploration of optimal cut scores for the Autism Index 6.

Research Question 8

When students with developmental disabilities, in a center-based special education setting, are rated by special education teaching staff using the GARS-3, how well does the GARS-3 discriminate individuals with ASD from individuals with other developmental disabilities that require substantial support?

Hypothesis 8a. It is predicted that the mean GARS-3 Autism Index 6 score for students with ASD will be significantly higher than the mean Autism Index 6 score for students with other developmental disabilities.

Means, significance values, and Cohen's *d* estimates were calculated for each GARS-3 subscale in addition to the two composite scores (i.e., Autism Index 4 and Autism Index 6). All mean comparisons were directional, given that the ASD group mean was expected to be significantly higher for all composites and subscales from an ASD screening instrument. For comparisons that involved a significant Levene's Test for Equality of Variances prior to the analysis, the analysis was corrected so that equal variances were not assumed. (These corrections are noted in Table 21.) Results were as follows: Restricted/Repetitive Behavior: $t(df)$ = 4.886(288), $p < 0.001$, d = 0.692; Social Interaction: $t(df)$ = 5.003(288), $p < 0.001$, d = 0.708; Social Communication: $t(df)$ = 5.045[a](84.478), $p < 0.001$, d = 0.833; Emotion Regulation: $t(df)$ = 5.761(288), $p < 0.001$, d = 0.816; Cognitive Style: $t(df)$ = 2.442(288), $p = 0.008$, d = 0.346; Maladaptive Speech: $t(df)$ = 4.004(288), $p < 0.001$, d = 0.567; Autism Index 4: $t(df)$ = 6.179[a](91.548), $p < 0.001$, d = 0.947; Autism Index 6: $t(df)$ = 6.384[a](88.401), $p < 0.001$, d = 1.009. All mean comparisons were statistically significant, though the comparison involving the Cognitive Style subscale would not be significant following an alpha correction for multiple comparisons. See Table 21 below for the subscale and composite score mean difference comparisons.

Table 21. Subscale and Composite Score Mean Difference Comparison Between Clinical Groups

Subscale/ Composite Scaled Score	Group	Mean	*SD*	*t* test	*df*	*p* value	Cohen's *d* (Cohen, 1988)[b]
Restricted/ Repetitive Behaviors	ASD	8.575	3.311	4.886	288	< 0.001*	0.692 (medium)
	Not ASD	6.313	3.121				
Social Interaction	ASD	9.297	3.447	5.003	288	< 0.001*	0.708 (medium)
	Not ASD	6.813	3.711				
Social Communication	ASD	9.730	2.935	5.045[a]	84.478	< 0.001*	0.833 (large)
	Not ASD	7.094	3.878				
Emotional Responses	ASD	9.031	3.281	5.761	288	< 0.001*	0.816 (large)
	Not ASD	6.359	3.253				
Cognitive Style	ASD	8.128	2.506	2.442	288	0.008*	0.346 (small)
	Not ASD	7.281	2.236				
Maladaptive Speech	ASD	8.965	2.856	4.004	288	< 0.001*	0.567 (medium)
	Not ASD	7.359	2.739				
Autism Index 4 Composite	ASD	94.443	17.802	6.179[a]	91.548	< 0.001*	0.947 (large)
	Not ASD	76.984	20.521				

Table 21 (cont'd)

Autism Index 6 Composite	ASD	91.128	17.419	6.384[a]	88.401	**< 0.001***	1.009 **(large)**
	Not ASD	72.641	21.234				

Note. ASD Group *n* = 226; Not ASD group *n* = 64

[a]Denotes corrected value when equal variances were not assumed due to a significant Levene's Test for Equality of Variances; [b]Interpretive ranges (Cohen, 1998)

In all, the hypothesis for research question 8a was supported as the mean GARS-3 Autism Index 6 score for students with ASD (*M* = 91.128) were significantly higher than the mean Autism Index 6 score for students with other developmental disabilities (*M* = 72.641; *t*(*df*) = 6.384[a] (88.401), *p* < 0.001**,** *d* = 1.009, large effect size).

Hypothesis 8b. Using the author-recommended cut score of 70 on the Autism Index 6, the sensitivity for accurately identifying risk level for cases with ASD will be ≥ .90.

Results indicated sensitivity of .854, which represented the true positives identified from the sample divided by the total number of "true" known ASD cases in the sample. Given .854 < .90, hypothesis 8b was not supported.

Hypothesis 8c. Using the author-recommended cut score of 70 on the Autism Index 6, the specificity for accurately identifying those not at risk for ASD will be ≥ .80.

Results did not support this hypothesis – yielding a specificity of .516. This indicated that the Autism Index 6 with this cut score accurately identified 51.6% of "true" known non-ASD cases as not at risk" for ASD. (See Table 22 for the binary classification matrix.)

Table 22. Study Three Binary Classification Matrix for Autism Index 6 with Cut Score of 70

GARS-3 Autism Index Score	Known Diagnosis ASD	 Non-ASD DDs	Sample Size
At risk (Autism Index ≥ 70)	193 (85.4%) *(True positives)*	31 (48.4%) *(False positives)*	*n* = 224
Not at risk (Autism Index ≤ 69)	33 (14.6%) *(False negatives)*	33 (51.6%) *(True negatives)*	*n* = 66
Sample Size	*n* = 226	*n* = 64	*N* = 290

Note. The percentage in each cell is the percentage per column.

The Pearson chi-square value for the binary classification matrix was 38.759 ($p < .001$) with an effect size *w* of .366 (medium effect; Cohen, 1988). Positive and negative predictive power were also calculated. The positive predictive value (i.e., true positives / [true positives + false positives]) was equal to 0.862 and indicates how likely that an individual in the sample with a positive screening decision (i.e., Autism Index 6 ≥ 70; Lalkhen & McCluskey, 2008) actually has an ASD diagnosis. The negative predictive value (i.e., true negatives / [true negatives + false negatives]) was equal to 0.500 and indicates how likely that an individual in the sample given a negative screen (i.e., Autism Index 6 ≤ 69; Lalkhen & McCluskey, 2008) does not have ASD.

Exploratory Analysis (8d). This exploratory analysis was conducted using Receiver Operating Characteristic (ROC) Curve analysis to examine the range of possible cut scores on the GARS-3 Autism Index 6 for purposes of determining the optimal screening cut score in the context of the study three sample.

The ROC area under the curve for the current cut score of 70 was calculated: AUC = 0.746, $p < .001$, CI (95%) = 0.674 – 0.819. (See Table 23 for cut scores and their associated levels of sensitivity and specificity.) When the cut score is lowered, the sensitivity increases, but specificity decreases. When the cut score is raised, sensitivity falls and specificity increases. At a lower cut score of 64, the sensitivity reached the value of .90, which would be consistent with the minimum target value from Hypothesis 8b. However, with this cut score specificity fell to 0.438. At a cut score between 92.5 and 93.5, specificity reached .80, which is consistent with the minimum specificity target for Hypothesis 8c. However, for these cut scores, sensitivity fell to between 0.491 and 0.535. When looking at a cut score of 85, there is a "balance point" where sensitivity roughly equals sensitivity at a value of 0.69 for both.

Table 23. ROC Curve Analysis: Sensitivity and Specificity by Cut Score

GARS-3 Interpretation Guide		ROC Curve Analysis		
Autism Index Score	Probability of ASD/ DSM-5 Severity Level	Autism Index Score	Sensitivity	Specificity
≤ 54	Unlikely	43	1.000	0.000
		45	0.987	0.063
		46.5	0.987	0.094
		48	0.982	0.109
		49.5	0.978	0.156
		51	0.978	0.203
		52.5	0.973	0.234
		54	0.973	0.266
55-70	Probable/ Level 1	55	0.965	0.281
		55.5	0.965	0.281
		57	0.960	0.313
		58.5	0.960	0.344
		60	0.942	0.375
		61.5	0.942	0.391
		62.5	0.938	0.406
		64	**0.907**	0.438
		65.5	0.885	0.438
		67	0.881	0.469
		68.5	0.863	0.469
		70[a]	0.854	0.516

Table 23 (cont'd)

71-100	Very Likely/ Level 2	71	0.845	0.531
		71.5	0.845	0.531
		73	0.836	0.547
		74.5	0.819	0.578
		76	0.805	0.578
		77.5	0.774	0.578
		79	0.770	0.594
		80.5	0.752	0.609
		82	0.712	0.656
		83.5	0.695	0.672
		85	**0.690**	**0.688**
		86.5	0.664	0.688
		88	0.642	0.703
		89.5	0.606	0.703
		91	0.580	0.719
		92.5	0.535	**0.766**
		93.5	0.491	**0.813**
		95	0.465	0.828
		96.5	0.420	0.859
		98	0.403	0.859
		99.5	0.367	0.859
		100	0.327	0.891
≥ 101	Very Likely/ Level 3	101	0.327	0.891
		102.5	0.283	0.922
		104	0.239	0.938
		105.5	0.212	0.953
		107	0.181	0.953
		108.5	0.150	0.969
		110	0.128	0.969
		111.5	0.102	0.969
		113	0.088	0.969
		114.5	0.066	0.969
		116	0.040	0.969
		118.5	0.022	0.969
		120.5	0.013	0.984
		121.5	0.009	1.000
		124.5	0.004	1.000
		128	0.000	1.000

Note. Values in the ROC table were produced by the ROC method in SPSS Version 26.

However, several important whole number (e.g., values defining the Autism Index score ranges) values were calculated manually and added to the table as SPSS, by convention, does not include them. Bolded and underlined sensitivity and/or specificity values are intended to highlight important cut score benchmarks indicated in the text.

[a]Cut score used to determine "very likely" ASD diagnosis in GARS-3 based on the GARS-3 manual (Gilliam, 2013)

Summary of Results

In conclusion, the eight research questions were answered using multiple samples across three studies. The tables below (see Table 24 through Table 26) provide a summary of the research questions, hypotheses, analyses/methods, and results to clarify the main points of the three studies prior to further, nuanced discussion.

Table 24. Study One Results Summary Table

	Research Question/ *Hypothesis (When Present)*	Method/*Analysis*	Results
1	How many **potentially interpretable factors** are present and should be **considered for retention**?	EFA with PAF *Scree plot, Kaiser criterion, Velicer's MAP, parallel analysis (PA)*	The more conservative/reliable criterion tests for factor retention (e.g., Velicer's MAP and PA) suggested retention ranging between **five and eight factors**.
2	How many **factors should be retained** to yield the most interpretable factor solution?	EFA interpretive procedure *Five researchers independently interpret factors across the range of likely solutions and retain the most interpretable factor solution by consensus*	The **six-factor model** yielded the most interpretable solution.

Table 24 (cont'd)

3	Are there **substantive correlations** between at least some GARS-3 factors within the most interpretable factor structure? *Correlations between at least some factors will be ≥ **.30**.*	EFA with oblique rotation *Examine the inter-factor correlation matrix to determine if substantive correlations are present*	**Five of the six factors were substantively correlated** with each other (i.e., ≥ .30), while one factor (Cognitive Disposition) had non-substantive correlations (near zero) with other factors.
4	How does the **six-factor solution correspond to the six GARS-3 subscales** proposed by the author (Gilliam, 2013)?	Qualitative comparison, calculation of the percentage of overlapping items per factor *Examine the factor constructs/names of the six-factor solution, compared to the six GARS-3 subscales*	The six factors were **highly consistent** with those identified by the instrument author (i.e., five of six factors involved 100% item overlap and one factor had 93.33% item overlap across the two models).

Note. Velicer's MAP = Velicer's minimum average partial correlation; EFA = Exploratory factor analysis; PAF = Principal axis factoring; GARS-3 = Gilliam Autism Rating Scale – Third Edition

Table 25. Study Two Results Summary Table

	Research Question/ *Hypothesis (When Present)*	Method/*Analysis*	Results
5	Does the interpretive **model proposed by the GARS-3 test author produce a reasonable fit** to the confirmatory sample inter-item covariance matrix? *The GARS-3 model will yield a reasonable fit to the confirmatory sample inter-item covariance matrix.*	CFA with WLSMV χ^2, *SRMR, RMSEA, CFI, TLI*	The study one model demonstrated **good fit** with the confirmatory sample.

Table 25 (cont'd)

6	Does the **retained factor solution from the study one EFA produce a reasonable fit** to the confirmatory sample inter-item covariance matrix? *The study one solution will reasonably fit the confirmatory sample inter-item covariance matrix.*	CFA with WLSMV χ^2, *SRMR, RMSEA, CFI, TLI*	The published GARS-3 model demonstrated **good fit** with the confirmatory sample.
7	When the GARS-3 author-proposed model and the EFA-generated model from study one are compared, **does one model show evidence of better fit** to the confirmatory sample inter-item covariance matrix? *The study one EFA model will show a substantively better fit with the inter-item covariance matrix of the confirmatory sample.*	CFA with WLSMV, MLR *Mplus DIFFTEST (adjusted* χ^2*) from WLSMV and/or AIC and BIC from MLR*	The two models **could not be directly compared** using the Mplus DIFFTEST because they differed only in the placement of one item. However, AIC and BIC values were lower for the **study one EFA model** suggesting **slightly better** fit. The Mplus DIFFTEST was conducted between the model from study one with and without cross-loadings. Results indicated the **model with cross-loadings had a better fit**.

Note. GARS-3 = Gilliam Autism Rating Scale – Third Edition; SRMR = Standardized root mean square residual; RMSEA = Root mean square error of approximation; CFI = Comparative fit index; TLI = Tucker-Lewis index; CFA = Confirmatory factor analysis; WLSMV = Weighted least squares mean variance; EFA = Exploratory factor analysis; AIC = Akaike information criterion; BIC = Bayesian information criterion; MLR = Robust maximum likelihood.

Table 26. Study Three Results Summary Table

	Research Question/ *Hypotheses*	Analysis/Method	Results
8	How well does the GARS-3 **discriminate** between individuals with **ASD** from individuals with **other developmental disabilities** that require substantial support?		
8a	*The **mean GARS-3 Autism Index 6 score for students with ASD will be significantly higher** than the mean Autism Index 6 score for students with other developmental disabilities.*	Mean differences, independent samples *t*-test, Cohen's (1988) effect size *d*	Autism Index 6 score for students with ASD (*M* = 91.128) were **significantly higher** than the mean Autism Index 6 score for students with other developmental disabilities (*M* = 72.641; *t*[*df*] = 6.384[a] [88.401], *p* < 0.001) and the standardized mean difference **(***d* = 1.009) was consistent with a **large effect size**.
8b	*The **sensitivity** for accurately identifying risk level for cases with ASD will be* ***≥ .90****.*	Sensitivity calculation	Results indicated a sensitivity of **.854**, which was less than the hypothesized level.
8c	*The **specificity** for accurately identifying those not at risk for ASD will be* ***≥ .80****.*	Specificity calculation	Results indicated a specificity of **.516**, which was less than the hypothesized level.
8d	*Exploratory analysis: Determination of the **optimal cut score** for the GARS-3 Autism Index 6.*	ROC Curve analysis to examine the range of potential cut scores.	**Sensitivity** achieved the hypothesized level (.90) at a cut score of **64**, but specificity suffered (.438). **Specificity** achieved the hypothesized level (.80) between cut scores of **92.5 and 93.5**, but sensitivity suffered (.535 - .491). **Sensitivity and specificity** were equal (approximately .69) at a cut score of **85**.

Note. GARS-3 = Gilliam Autism Rating Scale – Third Edition; ROC = Receiver Operating Characteristic.

CHAPTER FIVE: DISCUSSION

Taken together, the present series of studies examined important aspects of the construct validity of the Gilliam Autism Ratings Scale—Third Edition (GARS-3; Gilliam, 2013), in ASD and combined ASD and non-ASD developmental disabilities (DD) samples, when rated by special education staff members. The first two studies examined the internal structure validity of the GARS-3, utilizing both exploratory and confirmatory analyses in two separate samples. These studies provided the first independent examinations of the internal structure of the GARS-3—independent of the instrument author and using samples independent from the standardization sample or other samples reported in the manual. Beyond the internal structure of the instrument, a critical aspect of any clinical measure is how well it discriminates between clinical groups (i.e., level two screeners, such as the GARS-3, are intended to assess for more specific diagnoses after cases have been identified as "at-risk" [Kuriakose & Shalev, 2016; Norris & Lecavalier, 2010b]). The GARS-3 is a widely used level two ASD screener (e.g., among the most widely used by school psychologists for ASD assessment; Benson et al., 2019), but previously had limited independent research concerning its validity. Together, the series of three studies conducted for this dissertation, broke new ground in terms of independent research on the GARS-3, with some findings supporting aspects of GARS-3 validity and other findings raising concerns. These findings have at least tentative implications for practice, and, when taken together with study limitations, provide clear directions for future research.

In this section, results from Chapter Four will be discussed in detail. A broad summary of each of the three studies and their most salient findings will be reviewed. Specifically, for study one, model comparison and inter-factor correlations will be discussed. Discussion of study two will review model fit, as well as issues concerning item cross-loadings and other, potentially problematic items. The study three discussion will examine mean differences, sensitivity and specificity, and cut scores. Strengths and limitations of the studies will be addressed including those involving the sample and raters, generalizability of the results, and model testing. Finally, theoretical, practical, and research implications will be considered, with an emphasis on important considerations for, and contributions to, the field.

Study One: Exploratory Factor Analysis

Study one examined special education staff ratings of the GARS-3 for a sample of 204 students with ASD using exploratory factor analysis (EFA). Using principal axis factoring with a promax rotation, several indicators informed factor retention criteria (i.e., scree plot, Kaiser criterion, Velicer's MAP, and parallel analysis). Research question one sought to identify the number of factors considered for retention; using the aforementioned criteria, solutions ranged from three to nine factors. Emphasizing more reliable methods (i.e., Velicer's MAP and parallel analysis), a range of five to eight factors was suggested for retention. Factor solutions were independently examined for interpretability – as dictated in research question two – by five different evaluators (i.e., advanced doctoral students and faculty – familiar with the core and associated features of ASD and other developmental disorders). After discussion of different models, it was decided to retain a six-factor solution. Research question three examined correlations among the factors of the chosen solution, predicting that correlations between at least some factors would be ≥ .30. While five of six factors met this criteria, one factor (i.e., the fifth factor representing cognitive dispositions) had negative and near zero correlations (ranging from -0.29 to 0.14) with the other factors. Lastly, research question four compared study one's model to the published six-factor structure from the GARS-3 manual. These models were largely the same, with primary differences including one item fitting on a different factor and several substantive item cross-loadings. The following sections specifically discuss model comparison and inter-factor correlations of study one.

Model Comparison

Research question four sought to compare the factor solution from the current study to the published GARS-3 six-factor model (Gilliam, 2013). Both the study one EFA model's and published GARS-3 model's factor structures yielded a six-factor solution as the most interpretable solution. Further, the majority of items (57/58 or 98.28%) were assigned to the same factor across the two models. The single item from study one that loaded higher on a different factor compared to the published version was item 28 (i.e., responds inappropriately to humorous stimuli). This item is found on the factor Social Communication in the published model, while it loaded highest on the current study's factor, named Social-Emotional Reciprocity (named Social Interaction in the published model). While it loaded highest on a different factor in the study one model, the loading (.46) was not as substantive compared to the loadings of other items on that same factor. Additionally, it had a substantive cross-loading of .39 on the factor on which it was expected to load from the published model (i.e., Social Communication from the GARS-3 published model and Social Understanding in the current study EFA model). This finding suggests that this item may be substantively influenced, to a similar degree, by two different constructs related to social communication and interaction, which may make it difficult to interpret. If replicated in other samples (such as it was in the study two CFA), in subsequent revisions of the GARS, it would be beneficial to closely look at the wording or inclusion of this item.

In addition to item 28, there were seven other items that yielded substantive cross-loadings found in the present study. In the study one EFA, there were cross loadings on factors including Restricted & Repetitive Behaviors (i.e., cross loadings with Social-Emotional Reciprocity and Speech and Language), Cognitive Disposition (i.e., one cross-loading with Speech & Language), and Speech & Language (i.e., one cross-loading with Social Interaction). There were four cross loadings between the Social-Emotional Reciprocity and Social Understanding factors. The eight items that substantively loaded onto a factor it was not primarily assigned were: item 1: majority of alone time spent in repetitive or stereotyped behaviors; item 2: preoccupied with specific stimuli; item 29: difficulty understanding jokes; item 14: does not initiate conversations; item 28: responds inappropriately to humorous stimuli; item 55: repeats unintelligible sounds; and item 47: talks about the same things excessively. This finding is important to highlight because these items specifically should be considered for revision or removal from the rating scale. Because these items have more than one substantive loading pertaining to different factors, the factor model is not as clean or clear (Volker et al., 2016). Additionally, in the GARS-3 manual, there were no reported cross-loadings (Gilliam, 2013).

Given that most items loaded on the same factors across studies, the same constructs were identified in the evaluation of diagnostic and associated features of ASD. In the current study, a team of four ASD researchers, independently (both independently across these researchers and without reference to the subscale names used in the GARS-3 manual) named the construct represented by each factor based on item loadings and subsequently, came together discuss the factor names. The final factor names, converged upon after final discussion, were slightly different compared to the subscale names of the published GARS-3 (e.g., Social Interaction was interpreted as Social-Emotional Reciprocity; Social Communication was interpreted as Social Understanding). The researchers' process in naming factors emphasized item content while attempting to avoid naming the subscale in a positive or negative light. This approach was agreed upon, specifically thinking about how parents or raters (e.g., special education staff as in the current study) may view the names of the subscales upon rating – such that negative subscale names could be upsetting or discouraging). One example of this is the Maladaptive Speech factor from the published model was interpreted similarly but named Speech & Language (i.e., a more neutral factor name). Please refer to Table 17 for a more in-depth comparison of factor names and item loadings.

Inter-Factor Correlations

The hypothesis of research question three predicted substantive correlations between at least some of the factors greater than or equal to .30 (Nunnally & Bernstein [1994] as cited in Pett et al. [2003]). This was true for inter-correlations among five of the six factors. The correlations among those five factors were generally consistent with what would be expected from a scale that uses an overall score (i.e., supports the computation of a composite score) that includes items from all five factors. However, one of the six factors, factor 5 (Cognitive Disposition), did not yield any inter-factor correlations reaching the .30 benchmark. This factor had negative-to-near-zero correlations with the other factors (i.e., -.29 with Social-Emotional Reciprocity, -.11 with Restricted & Repetitive Behaviors, -.01 with Emotion Regulation, -.26 with Social Understanding, and .14 with Speech & Language). This suggests a number of potential interpretive problems for a subscale based on this factor and for its inclusion with the other five subscales in an overall composite score.

It is important to note that before completing items for factors (i.e., subscales) 5 and 6 (Cognitive Style/Disposition and Maladaptive Speech/Speech & Language), there is a question on the GARS-3 protocol asking if the individual being rated is “mute.” If the answer is yes, then the rater should not complete the next two subscales. On several of the completed GARS-3 protocols in the present project, this question was not answered and therefore it was not clear whether it was noted accordingly by the rater. (However, all participants were known to use at least a small number of words.) Of note, the qualifying question and minimum language requirements for completing the last two subscales is quite vague (i.e., “is the individual mute?”). In contrast, the manual gives a slightly more detailed description of minimum language requirements compared to the one question on the protocol, but it is very unlikely that a rater would read/have access to the manual upon completion of the scale. Based on what is directly on the protocol, raters may interpret “mute” differently than intended, or than could be measured by the last two subscales. “Mute” is likely interpreted as no spoken words, but there may have been cases where language skills were lower than what was required to accurately assess behavior for subscale items.

Despite following the directions on the protocol, the vagueness of the question may have led to the inclusion of individuals within insufficient language skills to meet the assumptions of the Cognitive Disposition factor. This particular factor includes items such as using precise speech, having an exceptional memory, and has superior knowledge in specific subjects. There is a clear prerequisite of expressive language skills needed for raters to observe these behaviors; in other words, a zero rating (i.e., “not at all like the individual”) could have indicated the absence of assumed language skills required for the behavior to be expressed. Thus, for cases that do not meet the assumed capacity required for the item, the behavior’s absence could be due to reasons other than the intended construct. This is one potential explanation for the negative-to-near-zero correlations between the Cognitive Disposition factor and the other factors.

Speech and language delays are present in many developmental disorders, and are commonly comorbid with ASD (Talbott et al., 2020; Veness, 2012). Within individuals with ASD, research suggests about 25-30% will not develop complex speech (Tager-Flusberg & Kasari, 2013; Tager-Flusberg et al., 2005; Talbott et al., 2020). Speech impairments are also present in children with ASD and intellectual disability (ID). Looking at children identified with ASD through the Autism and Developmental Disabilities Monitoring Network, research indicated that children with ID (i.e., IQ less than 70) were somewhat more likely to have a language impairment compared to individuals with an IQ greater than 70 (Maenner et al., 2014). Further, the literature shows that more severe ID, associated with greater impairment of skills, is more prevalent within the ASD and ID comorbidity compared to an ID diagnosis alone (Fombonne, 2002; Thurm et al., 2019).

In all, language delays seem to occur more frequently with ID within ASD, and ID occurs more frequently within ASD. This gives further justification for language-loaded items to be closely examined on instruments assessing ASD-related symptoms. In a study by Syriopoulou-Delli et al. (2018), researchers examined social skills via a questionnaire, looking specifically at the relationships between scores and variables including ID and language. Results indicated that within children with ID and ASD – compared to just children with ASD – they exhibited lower scores measuring constructs such as reciprocity, participation, and asking questions. Further, researchers examined lower scores for the nonverbal group with ASD and indicated findings could be explained both because verbal skills play a role in social skills, but also that items involved for measuring social skills had the prerequisite of language (Syriopoulou-Delli et al., 2018).

On factors that measure deficits related to ASD, such as social skills and communication, higher scores typically indicate a greater association with ASD – like with the GARS-3. However, when individuals do not have sufficient language skills to express these difficulties related to ASD, they have lower scores not because they do or do not have ASD, but because they lack sufficient language skills to be rated for these things. Consequently, this lack of and/or negative correlation pattern with this factor could be explained by the absence of sufficient verbal skills for some cases (i.e., the correlations involving this factor were not as expected due to construct irrelevant reasons, as the item responses would not necessarily have the same meaning for those cases whose language skills were too low). This interpretation was supported by an examination of completed record forms for some cases included in the dataset, which suggested the presence of only minimal expressive language skills.

Study Two: Confirmatory Factor Analysis

Study two utilized confirmatory factor analysis to examine model fit using GARS-3 ratings from special education staff in a sample of students with ASD or other developmental disorders with similar support needs (n = 200). Given the ordinal and non-normal nature of the item data, the Mplus WLSMV estimator was used. Indicators of fit included χ^2, SRMR, RMSEA, CFI, and TLI from WLSMV, while AIC and BIC indices were based on the Robust Maximum Likelihood estimator (MLR). Research question five examined the fit of the published GARS-3 model and results indicated good fit (χ^2 = 2323.440, df = 1580, p < .001; RMSEA = 0.049; SRMR = 0.081; CFI = 0.981; TLI = 0.980; MLR-based AIC = 26367.365 and BIC = 26990.747). Research question six looked at the fit of the model derived from the study one EFA which also showed evidence of good fit, as hypothesized (χ^2 = 2361.659, df = 1580, p < .001; RMSEA = 0.050; SRMR = 0.081; CFI = 0.980; TLI = 0.979; MLR-based AIC = 26338.297 and

BIC = 26961.679). Research question seven sought to determine which of the two models showed evidence for better fit. While the hypothesis predicted the EFA-generated model would show a substantively better fit, both models did not substantively differ and could not be directly compared using the Mplus inferential DIFFTEST (i.e., the models were not nested and involved the same *df*). However, it was possible to directly compare fit of the six-factor models with and without cross-loadings. It is important to note that the model that included the significant cross-loadings yielded a better fit (χ^2 = 112.474, *df* = 8, *p* < .001) according to the DIFFTEST.

Model Fit

Each research question assessed model fit with the CFA sample, which included both individuals with ASD and other non-ASD developmental disorders. This type of mixed sample likely represents the population being assessed by a level two autism screener which seeks to provide more diagnostic clarity in its specificity of an ASD diagnosis after a level one screener highlighted the need for further assessment (Norris & Lecavalier, 2010b). A level two screener, such as the GARS-3, would likely be used with a population showing atypical development (e.g., a level one screening tool identified this individual as needing further testing/observation) which would likely encompass a population like the study two sample that consists primarily of ASD cases but also includes some non-ASD cases with partially overlapping symptomatology and similar support needs. The GARS-3 standardization sample included individuals with ASD – specifically, 61.3% had a sole diagnosis of ASD and 37.75% had one or more comorbid conditions. The manual indicated that including cases with both sole and comorbid diagnoses would be more representative of current prevalent rates (Gilliam, 2013). While the variability of the diagnoses in the standardization sample was different than the non-ASD diagnoses included in this second study (e.g., 31.5% of the CFA sample was diagnosed with language disorders, Fragile X Syndrome, ADHD, etc.), it is important to note that both samples yielded almost identical models. Thus, it is encouraging that this research supports a good model fit for usage of the GARS-3 with a population that includes not just ASD cases, but at least some cases from a broader range of developmental disabilities.

Fit of Model with Cross-Loadings

The fit of the model from study one and the published GARS-3 model could not be directly compared using the Mplus DIFFTEST because they had identical degrees of freedom, and the models were not nested. The two models largely differed only in the placement of one item (i.e., item 28). However, the DIFFTEST method was used to compare the fit between the EFA-based six-factor model with and without cross-loadings. In study one, cross-loadings across factors highlighted items that may be influenced by, or represent, multiple constructs. In this comparison, the objective was to determine if utilizing these cross-loadings would improve or worsen the fit of the model. Results indicated that incorporating cross-loadings did yield a significantly better fitting model. However, the practicality of scoring an assessment in a manner that goes beyond each item to contribute to no more than one of the available subscales may be more complex than the better fit is worth. Having to add these eight items on multiple different scales may turn scoring into a more tedious task, cause more construct-irrelevant variance issues for subscale scores, and potentially deter users from utilizing the measure. It would likely be more beneficial, seeing that the six-factor structure is a good fit for samples that include both ASD and other DD diagnoses, to examine items that load on multiple factors for revision or elimination to improve the measure by attaining simple structure (i.e., each item loading substantively on no more than one of the factors).

Low Item Loading

When examining the parameter estimates for this study, it is notable that item 56 (e.g., uses “he” or “she” when referring to self) on factor 6 (i.e., Maladaptive Speech/Speech & Language) had very low parameter loadings on both models (i.e., .264 on the study one model and .265 on the published GARS-3 model). This loading contrasts with the substantive loading (i.e., .49) of the same item on this factor in the study one EFA. This could be due to sampling variation in the CFA sample, could suggest that this factor may not be the best place for this item, or suggest that consideration should be given to revising or excluding this item. The item’s disturbance term explained 93% of the variance, which suggests that this item is likely measuring primarily something other than the factor/construct to which it was assigned (i.e., speech and language) in the CFA. Diagnostically, this is an important item attending to a feature of ASD: pronominal reversals (APA, 2013; Zane, 2021). However, these indicators suggest that this item should be examined more closely in future factor analyses and, if the present finding replicates or this language-oriented item ends up loading more on a non-language factor, the item may need to be revised.

Study Three: Clinical Discriminant Validity

Study three examined the clinical discriminant validity of the GARS-3 – specifically, how well it was able to discriminate between ASD and non-ASD developmental disabilities. Research question eight examined mean differences between the clinical groups on the overall composite Autism Index scores, explored the sensitivity and specificity of the instrument, and attempted to determine optimal Autism Index 6 cut scores. Depending on the language level of the participant, the composite score of the GARS-3 either includes four subscales (i.e., the rater answered “yes” to the question on the protocol, “is the individual mute?” and yields the Autism

Index 4) or six subscales (i.e., the rater answered “no” to this same question and yields the Autism Index 6). The research questions related to sensitivity, specificity, and cut score focus only on the Autism Index 6 as this was the most appropriate composite score for the present sample.

Mean Differences

Study three examined mean differences between the ASD and non-ASD clinical groups. In five of six subscales and both the Autism Index composites, the ASD group had significantly higher means compared to the non-ASD group ($p < 0.001$). The Cognitive Style subscale still yielded significant differences between groups, but it would not have been significant following an alpha correction for multiple comparisons ($p = .008$). The largest effect sizes for these group differences appeared in the Social Communication and Emotional Responses subscales and in both composite scores (i.e., Autism Index 4 and Autism Index 6).

The GARS-3 manual does not provide mean subscale and composite values for the non-ASD group involved in the discriminant section. To examine differences between the current study and the manual, the ID group from the manual was identified as the clinical group most closely aligned with the third study’s non-ASD sample (e.g., having high support needs to attend an alternative special education setting; the non-ASD sample in study three with 32.81% with IQ scores lower than 70 and an average IQ of 70.88). When the current study’s non-ASD group is compared to the ID group from the GARS-3 manual, the current study’s non-ASD mean Autism Index scores were lower than the mean scores reported in the manual for the ID group. For example, the mean Autism Index 4 and Autism Index 6 scores for the current study’s non-ASD DDs group were 76.98 and 72.64, respectively. As reported in the manual, the mean Autism Index 4 and Autism Index 6 scores for the ID group were 89 and 87, respectively (Gilliam, 2013). Not only were average scores in the present study lower, but they also have larger mean differences between the ASD and the non-ASD comparison groups. In the present study, mean differences for composite scores between the ASD and non-ASD groups were approximately 17 and 18 points. This was a larger gap when compared to differences between the ID group and the normative sample as reported in the GARS-3 manual--which indicated a difference of 11 and 13 points (Gilliam, 2013). The larger differences between clinical groups in the current study compared to those in the manual could be due to differences in the comparison groups across the two studies (e.g., fewer than half of the current study's non-ASD group cases had an IQ below 70, while all cases in the ID group reported in the manual met criteria for ID, which involves a low IQ among the criteria). Additionally, mean scores in the current study's non-ASD group (i.e., Autism Index 4 = 76.98; Autism Index 6 = 72.64) were higher than other manual-reported clinical groups without ASD or ID (e.g., ADHD group with Autism Index 4 and Autism Index 6 scores of 61 and 55; group with speech language impairment (SLI) with scores of 62 and 59; Gilliam, 2013). Overall, the study three mean scores for the non-ASD group fell in an expected range; they were lower than the manual's diagnostic comparison group with ID and higher than other diagnostic groups including ADHD and SLI.

Sensitivity and Specificity

Sensitivity seeks to accurately identify individuals who have ASD—though a more thorough assessment would be required to make the actual diagnosis in a clinical setting. Using the recommended cut score of 70, results of the current study indicated that the sensitivity of the instrument was .854 for the Autism Index 6, which was below the hypothesized .90 standard but exceeding the .80 noted as acceptable (Kuriakose & Shalev, 2016). As a level two screening instrument, individuals have likely already been identified as having atypical development and the data is used to provide evidence for diagnostic clarity. Therefore, given this utility of the instrument, specificity in ASD likelihood is perhaps more meaningful to examine (Kuriakose & Shalev, 2016).

Specificity (i.e., percent of children correctly identified as not having ASD [i.e., true negative cases]) in the current study using the recommended cut score of 70 was .516, which did not meet the hypothesized and accepted standard of .80 (Kuriakose & Shalev, 2016). While this falls well below the acceptable standard, it should also be noted that overlapping symptomatology may make it increasingly difficult to distinguish between ASD and other developmental disabilities (Volker et al., 2016). These results suggest that in the present study samples, as rated by special education staff, the GARS-3, using the cut score recommended in the manual, is better able to identify risk for ASD and less able to rule out cases without ASD. This pattern is not optimal for a level two screener because of the large number of false positives (i.e., non-ASD cases that meet or exceed the cut score for ASD; Kuriakose & Shalev, 2016).

It is also notable that these results differ from the sensitivity and specificity levels reported in the GARS-3 manual. Gilliam (2013) examined the sensitivity and specificity of the Autism Index 4 and Autism Index 6, using the recommended cut score of 70, by contrasting those with ASD and those from various groups including individuals with ADHD; emotional disturbances; learning disabilities; speech-language impairment; and non-ASD disabled group including diagnoses of ID, deaf, blind, ADHD, ED/BD, LD, and physical/health impairment (Gilliam, 2013). The latter group seems most like the current study's comparison group (e.g., non-ASD developmental disabilities) which the manual reported for an Autism Index 6 a sensitivity of .96 and specificity of .84.

Other level two ASD screeners, such as the Social Responsiveness Scale – Second Edition (SRS-2; Constantino & Gruber, 2012) and Social Communication Questionnaire (SCQ; Rutter et al., 2003a), can also vary in sensitivity and specificity. The SRS-2 manual indicates a sensitivity and specificity of .92 from the School-Age Form using a comparison group of unaffected siblings (Bruni, 2014; Constantino & Gruber, 2012) – a similar comparison to the ASD and non-ASD groups in the GARS-3 manual. In contrast, an independent study examining the Adult Form yielded similar sensitivity (.86), but a much lower specificity (.60) in a sample comparing ASD and non-ASD patients at a psychiatric hospital (Mandell et al., 2012). Research on the previous version of the SRS-2 indicated difficulties distinguishing between ASD and other behavior disorders (Bruni, 2014) – in other words, problems with specificity when differentiating between disorders with similar symptomatology, that might appear when comparing ASD to typically developing samples. When examining the SCQ, the initial standardization study sample (Berument, 1999) examined sensitivity and specificity among different groups using the cut score of 15 – the suggested score based on ROC curve analyses. When comparing ASD with other diagnoses excluding mental retardation (i.e., MR; what is now referred to as intellectual disability), sensitivity was reported as .96 and specificity as .80. When just comparing ASD and MR, sensitivity of the SCQ remained the same (.96), but specificity worsened (.67). Directly comparing the ASD group with other diagnoses, including MR, sensitivity declined to .85 but specificity increased to .75 (Berument, 1999; Rutter et al., 2003a). Independent research using the SCQ to differentiate from other developmental disorders suggest using a lower cut score of 11 to optimize its sensitivity and specificity. One study by Wiggins and colleagues (2007) found that when utilizing the cut score of 11, the measure achieved maximum sensitivity and specificity of .89. The findings in the current study reflect the same difficulty in distinguishing between ASD and non-ASD DDs and indicate a slightly lower sensitivity (i.e., present study: .864) and a lower specificity (i.e., present study: .516) compared to these rating scales of similar function.

In all, the sensitivity was somewhat lower in the present study as compared to the manual's comparison of similar groups (i.e., present study: .864; manual: .96) and the specificity was significantly lower than reported in the manual (i.e., present study: .516; manual: .84). Differences in findings may be attributed to various factors including sample characteristics, rater types, and data collection methods. While the comparison groups between the current study and the manual were similar, they were not an exact match in terms of diagnoses. Additionally, the type of rater differed between the two samples with the current study having special education teaching staff ratings only, while the manual used data from a variety of rater types. Of note, special education staff are likely more familiar, compared to other types of raters, with characteristics of ASD and how ASD may differ from other disorders, leading to the expectation that these raters may be better able to discriminate between groups. Despite this expectation, sensitivity and specificity levels were lower than anticipated (based on estimates reported in the manual). Further, the current study's data was collected through paper and pencil forms while the standardization sample utilized a combination of online (93%) and paper and pencil (7%) data collection (Gilliam, 2013). These are some differences between the present study and the published GARS-3 that may have contributed to sensitivity and specificity findings.

Cut Scores

It is important to note that the cut score for the GARS-3 has changed across editions. The first edition recommended a cut score of 90 (Gilliam, 1995), the GARS-2 lowered its cut score to 85 (Gilliam, 2006), and the most updated GARS-3 recommended a cut score of 70 (Gilliam, 2013). Previous independent research was critical of the sensitivity and specificity of the measure, directly related to these cut scores. South et al. (2002) examined the GARS – using the cut score of 90 – with an ASD sample using parent raters; data indicated a low sensitivity rating of .48 with this sample. Using the same version of the GARS, Lecavalier (2005) examined sensitivity in an ASD sample using parent and teacher ratings. They reported the sensitivity as .38 in addition to the finding that the study's sample had a lower average composite score compared to the average composite scores reported by the manual. Further, Volker and colleagues (2016), sought to determine the sensitivity and specificity of the GARS-2 – with 85 being the recommended cut score – using ASD and non-ASD samples with special education teaching staff ratings. Their findings suggested a sensitivity level of .65 and specificity level of .81. See Table 27 for a direct comparison of these three studies.

Examination of these findings suggests sensitivity levels lower than predicted in independent samples and relatedly, high rates of false negatives for previous versions (e.g., 52% [South et al., 2002]; 62% [Lecavalier, 2005]; 35% [Volker et al., 2016]). With the first edition, the false negative rate was higher as compared to the Volker et al. (2016) study which examined the GARS-2 and had a lower cut score. Lowering the cut score would likely improve sensitivity as it would include more cases that are potentially missed. However then, specificity may suffer. With the third edition further lowering the cut score (i.e., to 70), it would likely address the problems associated with high rates of false negatives and increase its sensitivity; however, it may lead to classifying more individuals as false positives and therefore, specificity may have suffered. This pattern was observed with the current study with sensitivity reported as .864, but specificity much lower at .516. Singularly examining the numerical cut score is a point of interest, however, one must also consider changes between editions of the GARS: changes between the first and second editions were minimal, while there were large changes made to the GARS-3 (e.g., number of items, structure of the instrument, etc.).

Table 27. Sensitivity and Specificity of the GARS Across Versions in Independent Research

Article	GARS Version	Cut Score	Sample	Raters	Average Composite Score	Sensitivity	Specificity
South et al. (2002)	GARS	90	ASD	Parents	90.10	.48	---
Lecavalier (2005)	GARS	90	ASD	Parents and Teachers	85.80	.38	---
Volker et al. (2016)	GARS-2	85	ASD v. non-ASD	Special Education Teaching Staff	90.81[a]	.65	.81

Note. GARS = Gilliam Autism Rating Scale; ASD = Autism spectrum disorder

[a] Composite score for the ASD sample

The sensitivity and specificity of the current sample both fell below the hypothesized standards (i.e., .90 and .80) and previous findings reported in the GARS-3 manual. When looking to increase sensitivity or specificity for a given sample, researchers can consider increasing or decreasing the cut score. This is modeled in the ROC curve table (see Table 23).

Identifying a more ideal cut score may depend on the purpose for which the measure will be used. From a screening perspective, greater sensitivity would allow more children with atypical development to be identified (Norris & Lecavalier, 2010b). As a level two screener, the GARS-3 should also prioritize specificity compared to level one screeners which prioritize sensitivity (e.g., M-CHAT; Kuriakose & Shalev, 2016). As stated previously, specificity in a screener is important to determine next steps (e.g., comprehensive assessment for ASD).

In clinical settings, it may be more important for increased sensitivity or to minimize false negative cases. Clinicians do not want to overlook cases that may warrant a clinical diagnosis, which would require a more comprehensive assessment. However, it is not practical or necessary to provide full comprehensive assessments to all with atypical development identified via high sensitivity rates (hence, the importance of multi-modal assessment that includes a screening phase emphasizing both sensitivity and specificity).

In contrast, research settings might prioritize specificity (i.e., ruling out non-ASD cases) more than clinical settings, have high confidence that their sample cases are true positives for ASD, and contain very few, if any, false positive cases (i.e., non-ASD cases that met the ASD cutoff and were mistakenly screened in). Thus, in a research setting, a higher cut score may be used compared to a clinical setting, as researchers may accept a higher risk of false negatives in order to keep the false positive rate low.

Sensitivity and specificity – including utilization of different cut scores – can be examined with these things in mind: standards set by the field and/or the setting (e.g., clinical or research). When looking at the ROC curve analysis in study three, lowering the cut score to 64 would increase sensitivity to the .90 standard and leads to the inclusion of more cases as potential ASD cases. This may be more beneficial for clinical settings, where the concern is with minimizing false negative or missed ASD cases. However, this cut score also decreases the specificity to 0.438, and includes a high number of both false positives and negatives. Practically, this might lead to many comprehensive evaluations for ASD that were not necessarily warranted. In contrast, by increasing the cut score to between 92.5 and 93.5, specificity reaches the .80 standard, but sensitivity of the measure suffers (e.g., between 0.491 and 0.535). This increased cut score may be more ideal for research settings to assure that most, if not all, cases in the sample are accurately identified as having ASD. While the higher cut score might exclude some ASD cases (i.e., in this case, false negatives), the cost in a research setting is likely not as high as the cost associated in clinical practice. Looking at the balance point between sensitivity and specificity (i.e., when the two are roughly equal), considering them equally important, a cut score of 85 would yield levels of about .69. These levels still do not yield great confidence in diagnostic accuracy of a level two screener, providing evidence that this measure should not be used in isolation and as part of a multi-modal assessment.

Examining the mean Autism Index 6 score of the ASD group (91.13) compared to the mean score of the non-ASD group (72.64), it seems pertinent for researchers to be aware of the 70-cut score of the GARS-3 which indicates on the protocol, a "very likely" probability of ASD. Given that in practice, a clinician would likely be seeking to distinguish individuals with ASD from individuals with other developmental disabilities who may present with similar symptomatology, it is noteworthy that the average composite score for the non-ASD group was higher than the recommended cut score on the GARS-3 for ASD. Based on the current sample, it is likely beneficial to consider increasing the cut score from 70 to improve specificity when examining an individual who presents with concerns that could be explained by another developmental disorder. As previously mentioned, this could be addressed by looking at sensitivity and specificity standards as well as making a value judgment depending on the type of setting (e.g., clinical settings may want to prioritize sensitivity while research settings may want more stringent specificity).

Strengths

Important Research Contribution

The results of these studies provided much-needed validity information regarding the GARS-3, as independent research for the measure is very limited at this time. Specifically, the study one EFA and study two CFA are the first independent factor analyses of the GARS-3 and were performed on samples independent of the GARS-3 standardization sample. The only prior factor analysis of the GARS-3 was an EFA reported in the GARS-3 manual, which used the normative sample for the instrument (Gilliam, 2013).

Sample Characterization and Raters

Cases in the various samples were diagnosed by licensed clinicians using either DSM-5, earlier DSM-IV criteria, or special education eligibility criteria. As such, results of the study will generalize to individuals diagnosed using similar assessment measures in clinical practice and special education settings. In comparison, it was not clear if the GARS-3 normative sample contained cases that were clearly confirmed beyond reports of those who sent the data from their sites around the country (e.g., the manual only specified that cases had received an ASD diagnosis, lived in the US, and fit the age range). Further, many ratings from the normative sample were collected online (i.e., 93%; Gilliam, 2013). A notable strength of the present study was that the diagnoses for ASD cases were confirmed by clinical staff.

The present series of studies also included cognitive ability and language development data to better characterize the sample. Reported cognitive deviation quotient data provided a clear sense of the distribution of cognitive functioning of each sample. This made the average and range of cognitive abilities of the sample clearer for purposes of generalization. Cases were also screened out that did not meet minimum language or communication requirements—

consistent with guidance provided in the GARS-3 manual. The “mute” versus communicative expressive language determination was done by staff raters, who were very familiar with the language development of the student they were rating, which is generally consistent with what occurs in practice when utilizing the GARS-3 protocol.

When completing ratings of the GARS-3, a near 1:1 staff rater-to-student ratio was maintained. The center-based special education agency has many support staff available across its education units which, in most cases, allowed for each staff member to rate only one student each. This strategy occurred and was prioritized to support the independence assumption across cases within each data set. This procedure kept subgroups of student ratings from becoming nested within the same rater. Attention to this issue was a considerable of the present series of studies, as many other factor-analytic studies do not account for this issue.

Data and Sample Size

Efforts to minimize missing item responses were highly successful. Upon data booklet collection, strong quality control methods were utilized (e.g., forms were immediately checked for missing responses and promptly corrected). The amount of missing data was very infrequent (i.e., 0.21% missing items across all cases, equal to 25 total missing item values across all cases).

In both the EFA and CFA, samples met or exceeded minimum sample size requirements (based on tables provided by MacCallum et al., 1999, p. 93). The MacCallum et al. simulation study examined sample size adequacy for factor analyses, taking into account item communalities, anticipated number of factors and items per factor. With a six-factor structure, 58 items on the instrument, wide item communality estimates, and a sample size of 200 – as in both study one and study two – close to 100% of simulated solutions would be convergent or yield successful factor solutions (MacCallum et al., 1999; p. 93).

Factor Analytic Methodology

Strengths of the present research included the use of best practice reporting and statistical methods for EFA and CFA that were more rigorous than those reported in the GARS-3 manual (Basto & Pereira, 2012; Norris & Lecavalier, 2010a; Osborne & Banjanovic, 2016). In the manual, neither the inter-factor correlations nor the type of correlation input matrix used were reported. The present study emphasized the results of the inter-factor correlations and utilized procedures best suited for ordinal data (e.g., inter-item polychoric correlation matrix). The methodology from the manual reported using maximum likelihood estimation (i.e., a factoring procedure not robust to the violation of normality assumptions), while the present study used principal axis factoring given the anticipated non-normal data. Further, the GARS-3 manual reported using a scree plot and Kaiser criterion only for factor selection methods; the present study utilized both methods in addition to stronger, more accurate methods such as parallel analysis (Horn, 1965) and Velicer’s MAP (Basto & Pereira, 2012; O’Connor, 2000; Velicer, 1976; Velicer & Jackson, 1990). Similarly, the CFA involved best practice methodology for ordinal and non-normal data by using a diagonally WLS factor extraction procedure (Brown, 2015). While more rigorous methods were implemented in the present series of studies, the factor analytic results were generally similar to those reported in the manual, providing convergent evidence for the six-factor structure. More details about the strength of methodology are provided in the following paragraphs.

In the EFA, the use of the inter-item polychoric correlation matrix as the input accounted for the ordinal level of the item data. This is an important distinction as the Pearson correlation matrix tends to underestimate the degree of correlation with ordinal scales. The EFA also used principal axis factoring (PAF) which was appropriate for the anticipated non-normal nature of the data. This factoring procedure, compared maximum likelihood (ML) procedures, is more robust to deviations from normality in the data. This aspect is an important consideration for behavior data from ASD samples

The factor selection criteria and interpretive strategies in the EFA consisted of a range of high-quality indices. Gold-standard factor selection indices, including Velicer's MAP and parallel analysis (PA) were used to identify a range of interpretable solutions for further examination. The identified solutions were examined for interpretability by independent ASD experts who each selected the most interpretable solution and named the factors according to the constructs that appeared to be represented by the item loadings. The expert researchers all independently selected the same six-factor solution as most interpretable and proposed similar factor names and interpretations. Final factor names were determined through discussion and consensus. Thus, high quality indices informing factor retention to guide consensus on the most interpretable factor solution was a strength in the study's EFA methodology. Further, the use of multiple independent experts to determine the most interpretable solution and name the factors were also major strengths, as they increase confidence in the interpretation of factor analytic results.

The methodology within the CFA also utilized rigorous methodology that was best suited for the data. The ordinal and non-normal nature of the data were addressed by using the Weighted Least Squares Mean Variance (WLSMV) estimator in Mplus (Brown, 2015; Muthen & Muthen, 1998-2017). This estimator is ideal for categorical or ordinal data and is very robust to normality deviations (DiStefano & Morgan, 2014). Further, a variety of fit indices (e.g., χ^2, SRMR, RMSEA, CFI, TLI, etc.) were utilized as part of the CFA, which reflected different aspects of model fit. Beyond model fit indices, the Mplus DIFFTEST under the WLSMV estimator made it possible to directly compare nested factor models (e.g., the model with cross-loadings compared to the model without cross-loadings). For comparing non-nested models, information criterion indices (i.e., AIC and BIC) were used, based on Robust Maximum Likelihood (MLR) estimation.

Factor solutions were assessed with both an ASD only sample (i.e., in the EFA) and a mixed sample of individuals with ASD and other DDs (i.e., in the CFA). A six-factor model was a good fit for both samples. The samples were intentionally and systematically different, with the second more mixed sample having more variability. This mixed sample would be similar to students referred for an ASD diagnostic assessment, due to suspected ASD (e.g., with some having ASD and others having conditions with partially overlapping symptomatology). By increasing the variability in this way, it challenges the robustness of the EFA factor model established in a more homogenous sample. Despite the variability across the two samples, the six-factor solution fit well in the CFA sample.

Overall, the EFA, CFA, and comparison of nested models with and without cross-loadings, allowed for the identification of potentially problematic items. This included items with low primary loadings, items with substantive cross-loadings, cross-model comparisons which yielded a significantly better fit with the model including cross-loadings. This contribution is a core part of the recommendations for future studies, which will be further discussed in recommendations for future research section.

Clinical Discriminant Validity Sample and Analyses

Another strength of utilizing a mixed sample of both ASD and non-ASD DD cases was the ability to analyze mean comparisons between the two groups and utilize ROC curves to assess potentially useful cut scores for a variety of screening purposes. As mentioned, it is notable that the non-ASD DD sample represents a group that presents with many overlapping features with ASD and, as such, can be very difficult to discriminate from ASD when using an ASD screening or diagnostic assessment tool. This discriminant information is likely more beneficial information for clinicians as this sample better reflects the types of cases and distinctions practitioners need to make (e.g., discriminating between ASD and other DDs as opposed to discriminating ASD from neurotypical cases).

Limitations

Sample

While a strength of the series of studies was the relatively well characterized samples, there were some limitations in this area. The sample cases were diagnosed by licensed clinicians using clinical (e.g., DSM-5 or DSM-IV) or special education standards, but no gold standard assessment tool was consistently applied across cases in the sample. Ideally, such a measure is included in the diagnostic assessment to rigorously establish the diagnosis (e.g., ADOS-2, ADI-R; Kamp-Becker et al., 2021; Kuriakose & Shalev, 2016). While most individuals within this study likely completed such measures as part of the evaluation, it was not a requirement for services from the special education agency, and even when available, were not administered by a research reliable examiner. Though licensed clinicians determined diagnoses using DSM or special education criteria and administered a variety of assessment instruments, there was not a uniform battery of consistent assessment measures used across all cases in the samples.

To expand further on this issue of variability in assessment tools used across cases, the current study would clearly have benefitted from additional standardized, consistent assessment data to characterize the sample. Specifically, while data included the most recent cognitive assessment results for a high percentage of cases, there was no one consistent standardized cognitive assessment tool used across all cases. The samples were too diverse in terms of age, language abilities, etc. to use one standardized cognitive measure to adequately meet all of these needs. As such, it was reasonable – and more appropriate – to use the varied cognitive test data for sample description only and not inferentially in statistical analyses.

To supplement the cognitive profile, uniform adaptive behavior scores would have aided in the sample characterization, in addition to more complete information regarding comorbidities. Data on adaptive behavior is critical to establish a diagnosis of intellectual disability. While some information regarding comorbidities or non-ASD diagnoses was available, this information was not clearly available for all cases. Having this type of information uniformly across a whole sample would have been another way to better characterize the sample and make comparisons across potentially meaningful subgroups.

Additionally, standardized language scores would have been helpful to determine if the GARS-3 subscales with minimum verbal requirements should have been completed. However, individual standardized language scores were not available for a significant portion of students. Sufficient language for inclusion in this study was primarily determined from the question "is the individual mute?" as listed on the GARS-3 record form and use of any available spoken language data available through the agency. Because of the vague minimum language requirements on the GARS-3 for completing subscales 5 and 6, and lack of uniform language measures across cases, it was not possible to precisely characterize the samples in terms of language levels.

Raters

The studies would have benefited from more specific characterization of the raters. Overall, it was clearly known that all raters were special education teaching staff from a special education agency. The types of professionals were known (e.g., teachers, teacher assistants, teaching aides, speech pathologists, physical therapists, occupational therapists, etc.), but the precise number of each type of rater/profession or relative percentages for rater types was not available. Although it was noted that the majority of raters were special education teachers (e.g., an estimated 40% at the agency from 2008), the type and exact number of each subgroup of raters would be even more useful. Across staff members, there are likely to be different experiences and training which could lead to differences in how they completed ratings. This information could be used to potentially examine difference across rater subgroups. This is also a notable difference upon comparison with the normative sample of the GARS-3 which included approximately 23% parents and 58% teaching staff, with the other 19% reported to be a combination of other family members, school administrators, school counselors, consultants, advocates, behavior analysts, therapists, and case managers (Gilliam, 2013). Differences in the distribution of rater types in the present study compared to the normative sample could potentially be relevant to any notable differences in findings. Thus, having a more precise description of all raters could be very helpful in identifying potential moderators. However, the factor solutions found for the normative sample and the present EFA and CFA were convergent, supporting a significant degree of generalization across the different samples.

Generalizability

While this study contributes independent information regarding the validity of the GARS-3, it is important to note some of its limitations in generalizability. As noted above, the sample raters differ from the normative sample as reported in the GARS-3 manual (Gilliam, 2013; i.e., the normative sample had a variety of rater types while the present studies' raters consisted solely of those in special education teaching staff positions). Although this study found a very similar factor structure with the same number of factors and the same constructs (i.e., almost all items loaded on the same factors across models), it is not clear if any differences in the results of the current studies are generalizable to other/non-teaching raters (e.g., parents, mixed pool of raters, etc.).

Another large difference from the GARS-3 published model is that the second study assessed model fit using a mixed diagnostic sample (i.e., ASD and non-ASD DD). In the published model, the GARS-3 normative sample only included individuals with a diagnosis of ASD, though in many cases there was no evidence to suggest the diagnosis was confirmed as part of the data collection procedures. While this difference in samples could potentially lead to differences in models, it did not appear to do so across the EFA and CFA samples. In addition, several previous studies involving earlier editions of the GARS found largely the same or similar factor structures across more homogenous and more heterogenous samples, and across those involving different combinations of rater types (e.g., Lecavalier, 2005; Volker et al., 2022; Volker et al., 2016). Such convergent findings are good evidence for the generalizability of those findings.

It was noteworthy that several items were identified in the present studies that either showed substantive cross-loadings or in one case, loaded primarily on a different factor compared to the GARS-3 published model. The EFA in the GARS-3 manual did not suggest any substantive cross-loadings (Gilliam, 2013). It is possible that these differences could be the result of differences between samples and rater types; however, it is also important to note that the model reported in the manual was not clearly cross-validated on a second, independent sample (e.g., no CFA on a second sample). This left an open question about the degree to which any chance variation may have led to slight differences in results.

Model Testing

While this study is important in terms of providing much-needed independent research on the GARS-3, it was also limited in terms of the available model comparisons. (This is, at least partially, a function of there being no prior independent factor analytic studies of the GARS-3, and, therefore, no opportunities for prior studies to generate, explore, or suggest alternative factor models to test.) In the present study, only two models were available for testing in the CFA: (a) the published model from the GARS-3 manual (Gilliam, 2013) and (b) the model from the study one EFA. It is important to conduct additional independent factor analyses on other types of samples (e.g., different types of ASD samples, different age groups, those involving other DDs, etc.) and those using other types of raters. This could help better understand the generalizability of the known six-factor model and potentially suggest other models/variations of the known model.

Implications and Recommendations

Theoretical Implications

The internal structure of an assessment measure is important in terms of construct validity, especially in providing support for interpretation of the scores of the measure in terms of the constructs that they represent (Brown, 2015; Floyd & Widaman, 1995). In the first two studies, utilizing both exploratory and confirmatory methodology, underlying constructs partially informed by prior findings (i.e., the model) reported in the manual, and the degree of model fit were assessed (Brown, 2015; Floyd & Widaman, 1995; Gorsuch, 1983).

Broadly, the construct measured by the GARS-3 is ASD. The published factor names/measured constructs include Restrictive/Repetitive Behaviors, Social Interaction, Social Communication, Emotional Responses, Cognitive Style, and Maladaptive Speech (Gilliam, 2013). When examining DSM-5 diagnostic criteria for ASD, the first three GARS-3 subscales directly align with core diagnostic features (i.e., deficits in social communication and interaction and the presence of restricted and/or repetitive patterns of behavior, interests, or activities [APA, 2013]). Cognitive Style and Maladaptive Speech subscales appear to have items that both fit within or overlap with these two broad diagnostic categories, including how those diagnostic features present in different individuals (e.g., repeating words, talking about the same thing excessively). Items within the Emotional Responses subscale capture reactions to nonpreferred tasks or rigid thinking, which is related to core diagnostic features (e.g., restricted patterns of behavior), but might be characterized as an associated feature. Difficulties with emotion regulation are common – perhaps even more so – within the autism population (Cai et al., 2018); however, it is not an explicit diagnostic feature, nor is it specific to just those with ASD. While the six-factor model in the published GARS-3 was supported with near identical subscales/measured constructs in the results of the study one EFA, it is important to think about the differences between core and associated features that contribute to the overall measured construct, particularly when examining differences between factor models.

Taken together, results of the EFA yielding a factor structure consistent with the six proposed subscales of the published instrument and the CFA providing good support for the fit of this model, represent strong evidence of internal structure validity for the GARS-3 subscales. Because this structure was validated in the present context of special education staff ratings and Gilliam (2013) reported similar EFA findings in the context of ratings provided by a mixture of teachers, caregivers, etc., this set of findings provides initial evidence of construct generalization across rater types (Floyd & Widaman, 1995). However, it should be noted that internal structure validity is just one aspect of overall construct validity. In this case, the evidence clearly supports the relationships among the GARS-3 items being consistent with the proposed subscale scoring structure and, by extension, the intended constructs. However, criterion-related validity evidence is still required to support inferences to be made from subscale scores based on these factors (e.g., evidence of correlations between the GARS-3 scores and external measures of the same constructs, evidence of zero or low correlations with measures of constructs theoretically unrelated to the intended GARS-3 constructs, etc.).

Another important point of theoretical interest relates to the near-zero-to-negative inter-factor correlations involving factor 5. Preliminary evidence suggests that this unanticipated inter-factor correlation issue could be explained by the vague minimum language (i.e., not "mute") standard indicated in the GARS-3 manual (e.g., being too inclusive of those with more limited language skills). Skills beyond this minimum may be required in order to meet the communicative assumptions of some items, particularly those on the more language-loaded factors like factor 5. However, this explanation needs to be explored further in future research. Whether or not the minimum language requirement explains the lack of substantive inter-factor correlations involving factor 5, logically, lower-order factors should correlate positively with each other if their inter-correlations are to potentially yield a higher-order factor consistent with a composite score for the measure. If a lower-order factor is not correlated or is found to be orthogonal to other lower-order factors, it cannot be statistically connected to a higher-order construct. By extension, this means that a subscale based on this orthogonal factor would not have statistical support for inclusion in a composite score with the other subscales. Thus, it is critical that this issue with factor 5 be understood and appropriately addressed.

Implications for Practice

Given the use of the GARS-3 in school and clinical settings, this study provided important evidence from independent samples to further support validity and use with special education staff raters. The various editions of the GARS have been consistently among the top three ASD assessment measures used in schools (e.g., Aiello et al., 2017; Benson et al., 2019). Despite popular use of the GARS-3, there has been a scarcity of independent evidence of its psychometrics properties. Present EFA and CFA findings supported the internal structure validity of the GARS-3, supporting it proposed subscale structure—with subscales generally consistent, in terms of content, with the intended constructs. By examining the factor structure in both an ASD sample and a more mixed ASD and DD sample, the first two studies at least tentatively suggest some generalization of findings across the two types of samples. This is important because the more mixed sample is similar to the population on which the GARS-3 would be more typically used in practice (e.g., screening suspected cases, as part of a more comprehensive diagnostic evaluation, etc.).

Findings regarding the near-zero-to-negative inter-factor correlations involving factor 5 also have potentially important implications for clinicians and school psychologists who use the GARS-3 for screening or as part of a larger diagnostic evaluation. Though further research is needed regarding this issue, there is reasonable preliminary evidence to suggest that the lack of correlation between factor 5 and the other factors was due, in part, to the presence of cases in the samples with language skills that were lower than appropriate at least some of the items included in the factor. The manual gives very general guidance regarding sufficiency of language, or other communication strategies, required for subscales 5 and 6 to be completed (Gilliam, 2013; p. 10). Further, the protocol has one question (i.e., “is the individual mute?”) as the rule-in or rule-out item for completion of these two subscales. If the criterion is not met, then items for subscales 1 through 4 are completed, but not 5 and 6. The guidelines in the GARS-3 manual for minimum language and communication requirements were followed in recruiting cases for the present studies. Thus, it appears that the general, vague guidance provided in the manual concerning this issue may not be sufficient to assure that cases have sufficient language skills to be validly evaluated via some items from one or both of these factors. In the case of factor 5, several items assume sufficient communication skills for those being rated to convey their thoughts. This may have led to a pattern of item responses within the present samples where most cases were sufficiently verbal so that symptoms of ASD could be observed in their verbal/communicative behavior (i.e., higher item ratings were consistent with higher likelihood of ASD, as intended). However, some other cases were insufficiently verbal, which resulted in lower ratings on these items (or scores of 0), consistent with an associated language delay of ASD, but as a result, were also not indicative of the intended ASD symptoms that assume adequate language development. An examination of record forms and background characteristics indicated that this was a likely explanation for the factor 5 results. Though it is not statistically clear that factor 6 was impacted by this pattern, items on factor six do have minimum spoken language assumptions, in order to assess atypical use of speech and language (e.g., echolalia, pronoun misuse, etc.). Overall, based on these findings, it is recommended that practitioners who use the GARS-3 set a higher threshold than the GARS-3 manual suggests for the level of language development and communication required before items for subscales 5 and 6 are administered. Clearly, language-specific scales and/or criteria for completing these scales should be examined for possible revision to increase clarity of the constructs they are measuring; as a level two screener of ASD, these items should measure ASD, not language proficiency.

Additionally, results from study three indicated that, at least in the context of this type of sample with special education staff raters, a weakness of the measure is its specificity. Using the author-recommended cut score of 70 on the Autism Index 6, both sensitivity and specificity fell below hypothesized standards (i.e., observed .854 value compared to .90 hypothesis value for sensitivity; observed .516 compared to .80 hypothesis value for specificity). Based on results within the ROC curve table, study three sought to potentially determine an optimal screening cut score in this mixed sample of individuals with ASD and non-ASD DDs. However, there does not seem to be an ideal cut score for practitioners. Clinicians could consider using a cut score higher than the suggested 70 when trying to discriminate between different developmental disorders (e.g., cut score around 93), but, then the sensitivity levels drop below .50. The same issue exists when seeking to raise the sensitivity level to .90 – specificity drops to below .44. The balance point, or where sensitivity is about equal to specificity occurs around a cut score of 85 –higher than the recommended GARS-3 cut score. Sensitivity and specificity at this balance point are both approximately .69; still below the hypothesized levels for each. Overall, it was difficult to discern an optimal cut score with the current sample balancing both sensitivity and specificity. This measure has better sensitivity than it does specificity, which may lead to more false positives (e.g., individuals incorrectly being labeled as being “very likely” to have an ASD diagnosis). This is noteworthy as in a level two screener, the expected priority should be in specificity (Kuriakose & Shalev, 2016).

Practitioners should keep these issues in mind when considering use of the GARS-3 with these types of cases and these types of raters. The data does not suggest that it is reasonable, nor best practice, to use the GARS-3 in isolation, under these conditions, for diagnostic purposes. Best practice for diagnosis requires a comprehensive multi-source, multi-method assessment (e.g., Pandolfi & Magyar, 2016). Thus, if the GARS-3 were used, it should be one of several assessment strategies as part of a diagnostic assessment.

Research Implications and Recommendations for Future Research

A number of recommendations can be made for future research based on the findings and limitations of the present series of studies. This series of studies was the first to independently examine the factor structure and clinical discriminant validity of the GARS-3. This was completed using specific types of samples and raters. As such, many of the recommendations look to expand on and generalize these particular areas of validity. Given the limited independent psychometric research pertaining to the GARS-3, there is considerable need for additional studies across the various types of psychometric reliability and validity. Independent research using independent samples is critical for the evaluation of instruments, particularly those frequently used in practice and associated with higher stakes assessments. It is noteworthy, that independent research on prior editions of the GARS revealed poor sensitivity in a number of independent samples. Further, independent, replicated factor analyses identified a substantial number of items that loaded onto different factors compared to the then published model (e.g., Lecavalier, 2005; Volker et al., 2016; Pandolfi et al., 2010). In all, more independent research on the GARS-3 is needed to best understand its strengths and weaknesses, and to inform revision.

Future studies should also look at improving and widening the characterization of sample cases and raters to further support the validity of its use. Such future research would benefit by considering the current study's limitations of sample characterization in terms of diagnostic strategy (e.g., diagnoses based on DSM criteria vs. utilizing the ADOS-2/ADI-R in assessment methods), accounting for ASD comorbidities (e.g., ID, ADHD, seizure disorders, etc.), and use of a uniform or standard measure across all cases for critical domains such as cognitive ability, adaptive skills, and language development. As noted, demographic characteristics of the raters (e.g., age, gender, years of education, etc.) and their roles (e.g., caregiver, special education teacher, occupational therapist, one-to-one aide, speech/language pathologist, etc.) could be very helpful in better understanding potential rater-related moderating variables.

While the manual and current study examined the factor structure of the 58 total items, there is no research concerning the factor structure of the smaller subset of items that contribute to the Autism Index 4. Though intuitively, the four factors representing the four subscales that contribute to the Autism Index 4 would likely appear, as factors for that restricted set of items, when using the samples from the present studies, the Autism Index 4 is intended to be used primarily with cases where insufficient language and communication are present to allow for completion of the items for subscales 5 and 6. Thus, nonverbal cases could be included in the samples used to factor analyze the items for the Autism Index 4. This would include new types of cases that could alter findings and those new types of cases would be consistent with the intended use of the Autism Index 4.

The EFA and CFA in the present project yielded some aspects of the factor structure that differed from the published model/EFA findings reported in the GARS-3 manual. In the EFA reported in the manual, no items in the pattern matrix showed evidence of cross-loadings (i.e., items that loaded substantively on more than one factor; Gilliam, 2013). However, the current EFA and CFA yielded evidence of several items with potentially substantive cross-loadings. These items should be examined for cross-loadings in other samples that vary on potentially important characteristics (e.g., ASD severity, ASD and non-ASD cases, rater types, etc.). In the present studies, their cross-loadings appeared in the study one EFA sample and then, their inclusion improved the fit of the CFA model in the study two sample. Thus, the presence of cross-loadings in samples like those found in the present project rated by special education staff has already been cross-validated. Additionally, there was an item (i.e., item 56 related to misuse of pronouns) that had a substantive loading in the study one EFA (i.e., .49), but a low, non-substantive estimate within the CFA. This may be another item that fits better on a different factor but in any case, should be examined in future evaluations. In all, these items should be considered for revision, as substantive cross-loadings and inappropriate item placement can lead to the presence of construct irrelevant variance that could dilute or otherwise adversely impact interpretation in at least some cases.

As mentioned earlier, inter-factor correlation results showed that factor five (i.e., Cognitive Disposition/Cognitive Style) had near-zero-to-negative correlations with all other factors. This is clearly cause for concern, because this pattern is not consistent with including items from factor 5 in an overall composite score for the GARS-3. Therefore, it is important to better understand why this finding occurred. Evidence was found in the present study that insufficient language and communication development in some cases within the sample may have contributed to this finding—despite all cases meeting the GARS-3 manual's guidance regarding the communication minimum for completion of all items. This suggests the need to revisit the minimum language and communication requirements and provide clearer guidance. However, this more general issue with inter-factor correlations should be examined in other samples. The language hypothesis, among other potential explanations, could be explicitly tested with comparison groups, varying inclusion criteria across multiple analyses, etc.

The third study in this project included samples of different disorders that have partially overlapping features and as such, are more difficult to distinguish from each other. This represents a particularly difficult between-group discrimination. However, as these samples likely represent, to a considerable degree, the types of cases targeted for further assessment (e.g., ASD and other DDs) during screening, this is an important discrimination for an instrument, such as the GARS-3, to make. Future ROC curve analyses in other types of samples with other types of raters would be beneficial to further explore the most beneficial cut scores, especially for this type of difficult diagnostic discrimination that likely reflects real world clinical or special education use of the tool. It would also be helpful to better understand how well the instrument discriminates between other types of comparison groups (e.g., particular DDs, samples differing in level of cognitive impairment, etc.).

Finally, future research should examine the criterion-related validity of factors, or the subscales derived from them, with other standardized assessments measuring similar constructs (i.e., convergent validity) and those measures theoretically unrelated constructs (i.e., divergent validity). The GARS-3 manual reports large to very large correlations of the Autism Index 4 and Autism Index 6 scores with composite scores from the Autism Behavior Checklist (ABC), Childhood Autism Rating Scale – Second Edition (CARS-2), Gilliam Asperger's Disorder Scale (GADS), and Autism Diagnostic Observation Schedule (ADOS). This is good evidence for the criterion-related validity of the GARS-3 composite scores; however, more extensive criterion-related evidence is needed for the subscales whose internal structure validity have been supported by factor-analytic findings.

APPENDICES

APPENDIX A: GARS-3 Subscales and Composites

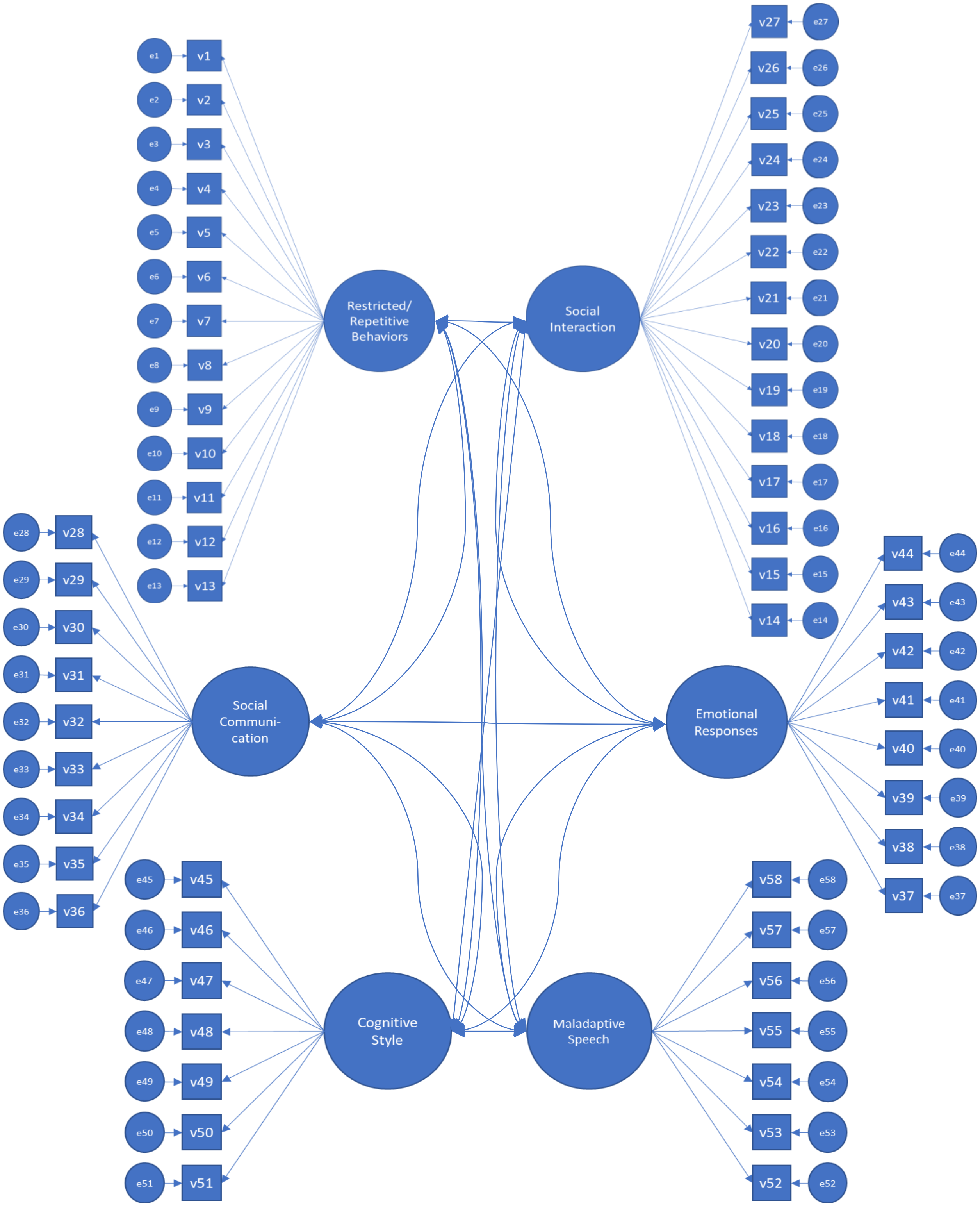
Table A1. GARS-3 Subscales

Subscale	Number of Items	Example of Subscale Content
Restricted/Repetitive Behaviors	13	Flicks fingers in front of eyes, stares at hand or other things for 5 seconds or more, unusual sensory interests
Social Interaction	14	Initiates conversation, expresses pleasure in interactions, shows interest in others
Social Communication	9	Understands joking, predicts consequences of social events
Emotional Responses	8	Needs excessive reassurance, extreme reactions to loud and unexpected noises
Cognitive Style	7	Precise speech, conversation reflects restricted interests
Maladaptive Speech	7	Presence of echolalia, flat tone, or affect, use of idiosyncratic words

Table A2. GARS-3 Composite Scores

Composite	Composite Subscales	Number of Subscales	Number of Items
Autism Index 4	Restricted/Repetitive Behaviors, Social Interaction, Social Communication, Emotional Responses	4	44
Autism Index 6	Restricted/Repetitive Behaviors, Social Interaction, Social Communication, Emotional Responses, Cognitive Style, Maladaptive Speech	6	58

APPENDIX B: Published GARS-3 (Gilliam, 2013) Six-Factor Model


Restricted/ Repetitive Behaviors
Social Interaction
Social Communi-cation
Emotional Responses
Cognitive Style
Maladaptive Speech
v1 v2 v3 v4 v5 v6 v7 v8 v9 v10 v11 v12 v13
e1 e2 e3 e4 e5 e6 e7 e8 e9 e10 e11 e12 e13
v14 v15 v16 v17 v18 v19 v20 v21 v22 v23 v24 v25 v26 v27
e14 e15 e16 e17 e18 e19 e20 e21 e22 e23 e24 e25 e26 e27
v28 v29 v30 v31 v32 v33 v34 v35 v36
e28 e29 e30 e31 e32 e33 e34 e35 e36
v37 v38 v39 v40 v41 v42 v43 v44
e37 e38 e39 e40 e41 e42 e43 e44
v45 v46 v47 v48 v49 v50 v51
e45 e46 e47 e48 e49 e50 e51
v52 v53 v54 v55 v56 v57 v58
e52 e53 e54 e55 e56 e57 e58

APPENDIX C: Study One Seven-Factor Solution Pattern Matrix

Item	Item Stem	Factor						
		1	2	3	4	5	6	7
19	Shows minimal expressed pleasure in interactions	**.95**	-.02	-.01	.05	-.03	.05	-.05
23	Shows minimal or no response when others attempt to interact	**.94**	.09	-.13	.01	-.07	-.03	.12
27	Shows little or no interest in others	**.93**	.03	-.07	.00	-.03	.01	.10
18	Seems indifferent to other person's attention	**.90**	-.05	.10	.01	-.17	.06	.02
22	Seems unwilling to get others to interact	**.88**	-.02	.04	.05	.02	.01	.08
20	Displays little or no excitement in showing toys or objects	**.85**	.04	.12	-.01	-.01	-.01	-.09
21	Seems uninterested in pointing out things	**.81**	-.06	.13	-.07	.12	.00	-.08
25	Doesn't try to make friends	**.81**	.02	.08	-.04	.17	-.01	-.04
15	Pays little or no attention to peers	**.77**	-.06	.08	.00	-.02	.08	.17
24	Displays little or no reciprocal communication	**.72**	.06	.06	-.13	-.01	-.03	.09
16	Fails to imitate	**.67**	-.01	.14	-.04	-.11	-.01	.26
14	Does not initiate conversations	**.63**	.02	.36	-.04	-.02	-.03	-.07
26	Fails to engage in creative play	**.61**	-.02	.27	-.15	.12	-.09	-.11
17	Doesn't follow other's gestures to look at something	**.59**	-.02	.18	-.02	-.05	-.01	.26
43	Tantrums when doesn't get their way	-.17	**.98**	.04	-.11	-.06	.06	.02
39	Tantrums when frustrated	-.05	**.95**	-.07	-.08	.03	.00	.09
44	Tantrums when told to stop something they enjoy	-.10	**.92**	-.04	-.16	-.04	.11	.06
41	Responds negatively when given commands	-.04	**.91**	.02	-.01	-.12	-.11	.10
40	Upset when routines are changed	.07	**.84**	-.01	.12	.07	.01	-.03
38	Frustrated when they cannot do something	.23	**.83**	-.13	.18	.01	-.06	.01
37	Needs an excessive amount of reassurance if things are changed	.07	**.82**	.20	.24	.03	-.06	-.17
42	Has extreme reactions in response to loud, unexpected noise	.14	**.63**	-.03	-.14	.07	.07	-.01

33	Difficulty understanding why people dislike them	.15	-.08	**.87**	.07	.06	-.06	.07
30	Difficulty understanding slang	.21	.05	**.82**	.00	-.08	-.02	.08
34	Fails to predict social consequences	.20	.07	**.82**	.06	.02	.06	-.10
31	Difficulty identifying teasing	.22	-.08	**.80**	.00	-.01	.04	.10
32	Difficulty understanding when being ridiculed	.22	-.16	**.80**	.00	.03	.00	.12
36	Doesn't understand that the other person doesn't know	.16	.08	**.76**	.00	.04	.13	-.07
29	Difficulty understanding jokes	.31	.08	**.69**	-.10	-.03	.01	-.03
35	Doesn't seem to understand people have different thoughts and feelings	.17	.15	**.69**	-.08	.07	.16	-.07
28	Responds inappropriately to humorous stimuli	.42	-.02	**.44**	.01	-.14	-.03	.25
48	Superior knowledge in specific subjects	.05	-.05	-.11	**.92**	.14	-.02	-.10
50	Intense, obsessive interest in specific subjects	.17	.04	-.14	**.92**	.04	-.08	-.03
49	Excellent memory	-.14	.04	.07	**.82**	.15	.03	-.17
46	Concrete meanings to words	-.03	.10	.21	**.80**	-.17	-.08	.07
45	Exceptionally precise speech	-.09	-.18	-.08	**.73**	-.12	.13	.10
51	Makes naïve remarks	-.18	-.05	.06	**.73**	-.11	.06	.19
47	Talks about same thing excessively	-.09	.15	-.03	**.62**	-.01	.35	.14
6	Flap hands or fingers	-.12	-.09	.01	.00	**.97**	.00	.03
4	Flicks fingers rapidly in front of eyes	-.09	-.05	.02	-.02	**.81**	.03	.18
5	Makes rapid lunging, darting movements	-.03	.15	.06	-.04	**.70**	-.09	.13
7	Makes high-pitched sounds or other vocalizations	.21	.02	-.07	-.26	**.55**	-.02	.24
3	Stares at hands, objects, or items in environment	.22	-.03	-.08	.01	**.47**	.06	.42
13	Ritualistic or compulsive behaviors	-.05	.22	.27	.18	**.39**	-.05	.37
53	Repeats words out of context	.06	-.03	-.05	.12	-.03	**.88**	.01
52	Repeats words or phrases	.01	-.11	.08	.06	.08	**.84**	-.06
58	Utters idiosyncratic words or phrases	.00	.08	.01	-.10	-.11	**.80**	.18

54	Speaks with flat tone, affect	.26	.07	.04	.13	.14	**.74**	-.40
57	Abnormal speech (tone, volume, rate)	-.04	.07	.17	.00	-.08	**.70**	.08
56	Uses “he” or “she” when referring to self	-.16	-.03	-.02	.27	-.15	**.45**	.28
55	Uses “yes” and “no” inappropriately	-.04	-.10	.35	-.09	.14	**.43**	.00
8	Uses toys or objects inappropriately	.04	.07	.22	-.05	.13	-.10	**.66**
10	Engages in stereotyped behaviors in play	.04	.03	.06	.04	.34	.02	**.66**
1	Majority of time alone spent in repetitive or stereotyped behaviors	.37	-.05	-.09	.21	.21	-.02	**.63**
2	Preoccupied with specific stimuli	.40	.09	-.25	.07	.17	.09	**.61**
12	Shows unusual interest in sensory aspects	.09	.04	.14	-.11	.18	.06	**.60**
9	Does things repetitively	.05	.15	.25	.25	.33	-.07	**.48**
11	Repeats unintelligible sounds	.08	.12	-.05	-.29	.21	.35	**.42**

Note. Loadings assigned to each individual factor are bolded. Loadings greater than .30 are underlined.

APPENDIX D: CFA of Published GARS-3 Six-Factor Model

Table D1. Study Two CFA of Published GARS-3 Six-Factor Model

Factor	Item	Item Stem	Parameter Estimate	Standard Error	*t* Statistic	Two-tailed *p*-value	R^2	Residual Variance
Social Interaction	14	Does not initiate conversations	0.836	0.031	27.174	< 0.001	0.699	0.301
	15	Pays little or no attention to peers	0.894	0.019	46.219	< 0.001	0.800	0.200
	16	Fails to imitate	0.859	0.023	36.660	< 0.001	0.738	0.262
	17	Doesn't follow other's gestures to look at something	0.844	0.024	34.841	< 0.001	0.712	0.288
	18	Seems indifferent to other person's attention	0.861	0.023	38.146	< 0.001	0.741	0.259
	19	Shows minimal expressed pleasure in interactions	0.898	0.017	52.511	< 0.001	0.806	0.194
	20	Displays little or no excitement in showing toys or objects	0.900	0.019	47.142	< 0.001	0.810	0.190
	21	Seems uninterested in pointing out things	0.954	0.013	74.066	< 0.001	0.911	0.089
	22	Seems unwilling to get others to interact	0.918	0.016	59.099	< 0.001	0.842	0.158
	23	Shows minimal or no response when others attempt to interact	0.901	0.017	53.598	< 0.001	0.813	0.187
	24	Displays little or no reciprocal communication	0.838	0.027	30.676	< 0.001	0.702	0.298
	25	Doesn't try to make friends	0.935	0.017	56.396	< 0.001	0.873	0.127

Table D1 (cont'd)

	26	Fails to engage in creative play	0.844	0.028	29.747	< 0.001	0.712	0.288
	27	Shows little or no interest in others	0.929	0.014	65.963	< 0.001	0.862	0.138
Restricted/ Repetitive Behaviors	1	Majority of time alone spent in repetitive or stereotyped behaviors	0.873	0.030	29.457	< 0.001	0.762	0.238
	2	Preoccupied with specific stimuli	0.931	0.018	51.856	< 0.001	0.867	0.133
	3	Stares at hands, objects, or items in environment	0.835	0.032	26.084	< 0.001	0.697	0.303
	4	Flicks fingers rapidly in front of eyes	0.702	0.052	13.607	< 0.001	0.493	0.507
	5	Makes rapid lunging, darting movements	0.782	0.044	17.843	< 0.001	0.612	0.388
	6	Flap hands or fingers	0.697	0.049	14.252	< 0.001	0.485	0.515
	7	Makes high-pitched sounds or other vocalizations	0.774	0.040	19.593	< 0.001	0.599	0.401
	8	Uses toys or objects inappropriately	0.869	0.029	30.347	< 0.001	0.754	0.246
	9	Does things repetitively	0.840	0.031	27.119	< 0.001	0.705	0.295
	10	Engages in stereotyped behaviors in play	0.875	0.028	31.782	< 0.001	0.766	0.234
	11	Repeats unintelligible sounds	0.797	0.036	22.065	< 0.001	0.635	0.365
	12	Shows unusual interest in sensory aspects	0.869	0.027	32.328	< 0.001	0.756	0.244

Table D1 (cont'd)

	13	Ritualistic or compulsive behaviors	0.890	0.029	31.054	< 0.001	0.792	0.208
Social Communication	28	Responds inappropriately to humorous stimuli	0.835	0.034	24.844	< 0.001	0.698	0.302
	29	Difficulty understanding jokes	0.953	0.013	72.854	< 0.001	0.908	0.092
	30	Difficulty understanding slang	0.947	0.013	75.465	< 0.001	0.896	0.104
	31	Difficulty identifying teasing	0.975	0.007	143.262	< 0.001	0.951	0.049
	32	Difficulty understanding when being ridiculed	0.990	0.004	220.460	< 0.001	0.979	0.021
	33	Difficulty understanding why people dislike them	0.981	0.009	110.002	< 0.001	0.962	0.038
	34	Fails to predict social consequences	0.957	0.012	77.558	< 0.001	0.916	0.084
	35	Doesn't seem to understand people have different thoughts and feelings	0.966	0.010	96.129	< 0.001	0.933	0.067
	36	Doesn't understand that the other person doesn't know	0.969	0.009	103.660	< 0.001	0.939	0.061
Emotional Responses	37	Needs an excessive amount of reassurance if things are changed	0.751	0.043	17.352	< 0.001	0.565	0.435
	38	Frustrated when they cannot do something	0.873	0.026	34.109	< 0.001	0.762	0.238
	39	Tantrums when frustrated	0.930	0.013	71.720	< 0.001	0.865	0.135
	40	Upset when routines are changed	0.901	0.024	36.953	< 0.001	0.812	0.188

Table D1 (cont'd)

	41	Responds negatively when given commands	0.896	0.019	47.551	< 0.001	0.802	0.198
	42	Has extreme reactions in response to loud, unexpected noise	0.775	0.047	16.500	< 0.001	0.601	0.399
	43	Tantrums when doesn't get their way	0.985	0.007	149.655	< 0.001	0.970	0.030
	44	Tantrums when told to stop something they enjoy	0.954	0.010	99.394	< 0.001	0.911	0.089
Cognitive Style	45	Exceptionally precise speech	0.713	0.047	15.253	< 0.001	0.509	0.491
	46	Concrete meanings to words	0.711	0.044	16.332	< 0.001	0.506	0.494
	47	Talks about same thing excessively	0.834	0.035	23.645	< 0.001	0.695	0.305
	48	Superior knowledge in specific subjects	0.864	0.036	24.089	< 0.001	0.747	0.253
	49	Excellent memory	0.690	0.047	14.693	< 0.001	0.476	0.524
	50	Intense, obsessive interest in specific subjects	0.867	0.029	29.955	< 0.001	0.751	0.249
	51	Makes naïve remarks	0.823	0.037	22.294	< 0.001	0.677	0.323
Maladaptive Speech	52	Repeats words or phrases	0.877	0.036	24.122	< 0.001	0.769	0.231
	53	Repeats words out of context	0.813	0.034	23.832	< 0.001	0.661	0.339
	54	Speaks with flat tone, affect	0.711	0.060	11.941	< 0.001	0.505	0.495

Table D1 (cont'd)

55	Uses "yes" and "no" inappropriately	0.735	0.059	12.476	< 0.001	0.540	0.460
56	Uses "he" or "she" when referring to self	0.265	0.114	2.321	0.020	0.070	0.930
57	Abnormal speech (tone, volume, rate)	0.811	0.051	15.900	< 0.001	0.658	0.342
58	Utters idiosyncratic words or phrases	0.888	0.040	22.341	< 0.001	0.789	0.211

Table D2. Study Two Inter-Factor Correlation Matrix: Published GARS-3 Six-Factor Model

Factor Number and Name	Factor Number					
	1	2	3	4	5	6
Factor 1: Restricted/Repetitive Behaviors	1.00					
Factor 2: Social Interaction	**.71**	1.00				
Factor 3: Social Communication	**.72**	**.83**	1.00			
Factor 4: Emotional Responses	**.61**	**.50**	**.53**	1.00		
Factor 5: Cognitive Style	.06	-.12	-.02	.12	1.00	
Factor 6: Maladaptive Speech	**.62**	**.49**	**.64**	**.47**	.26	1.00

Note. Correlations greater than or equal to .30 are bolded.

Figure D1. Path Diagram for Published GARS-3 Social Interaction Factor

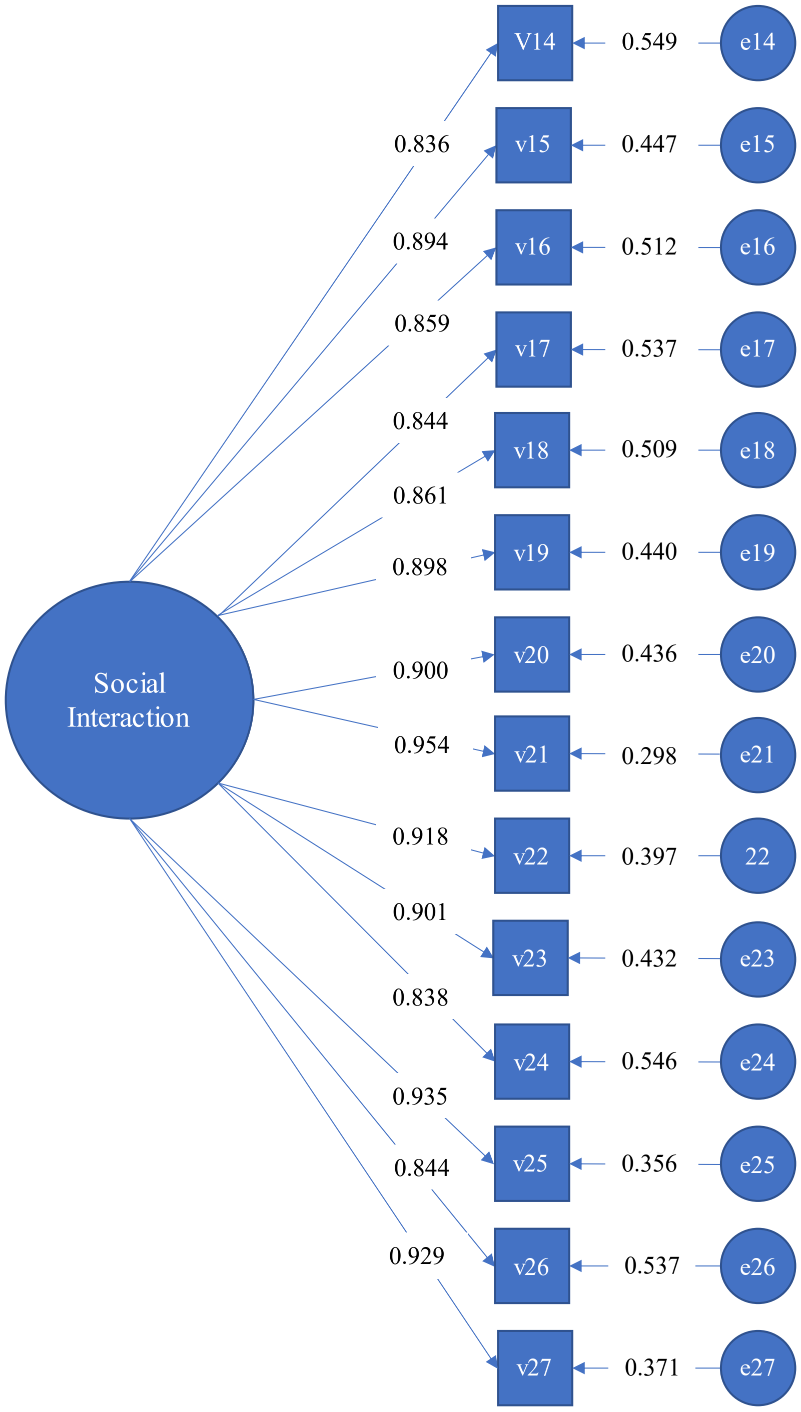


Figure D2. Path Diagram for Published GARS-3 Restricted/Repetitive Behavior Factor

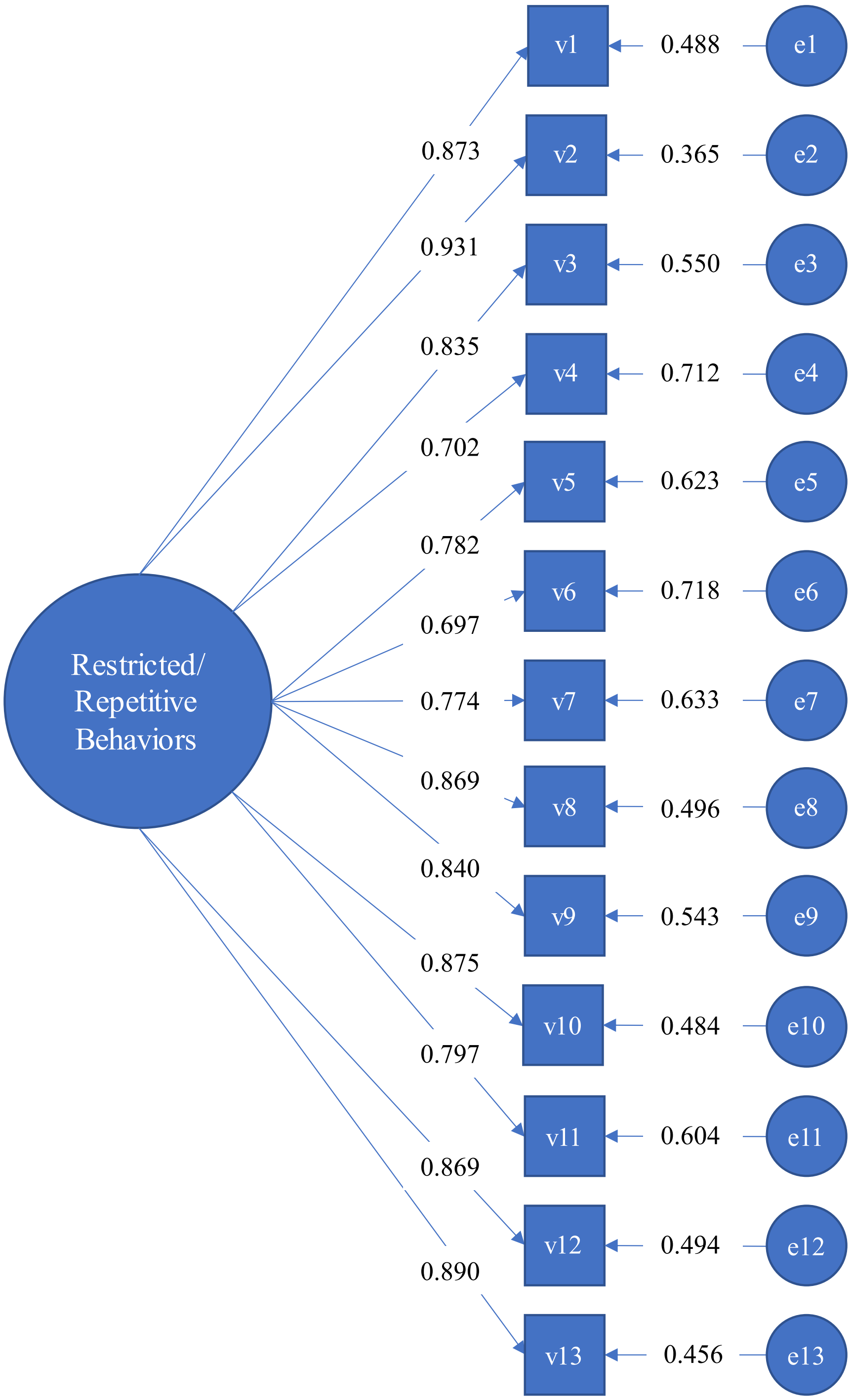


Figure D3. Path Diagram for Published GARS-3 Social Communication Factor

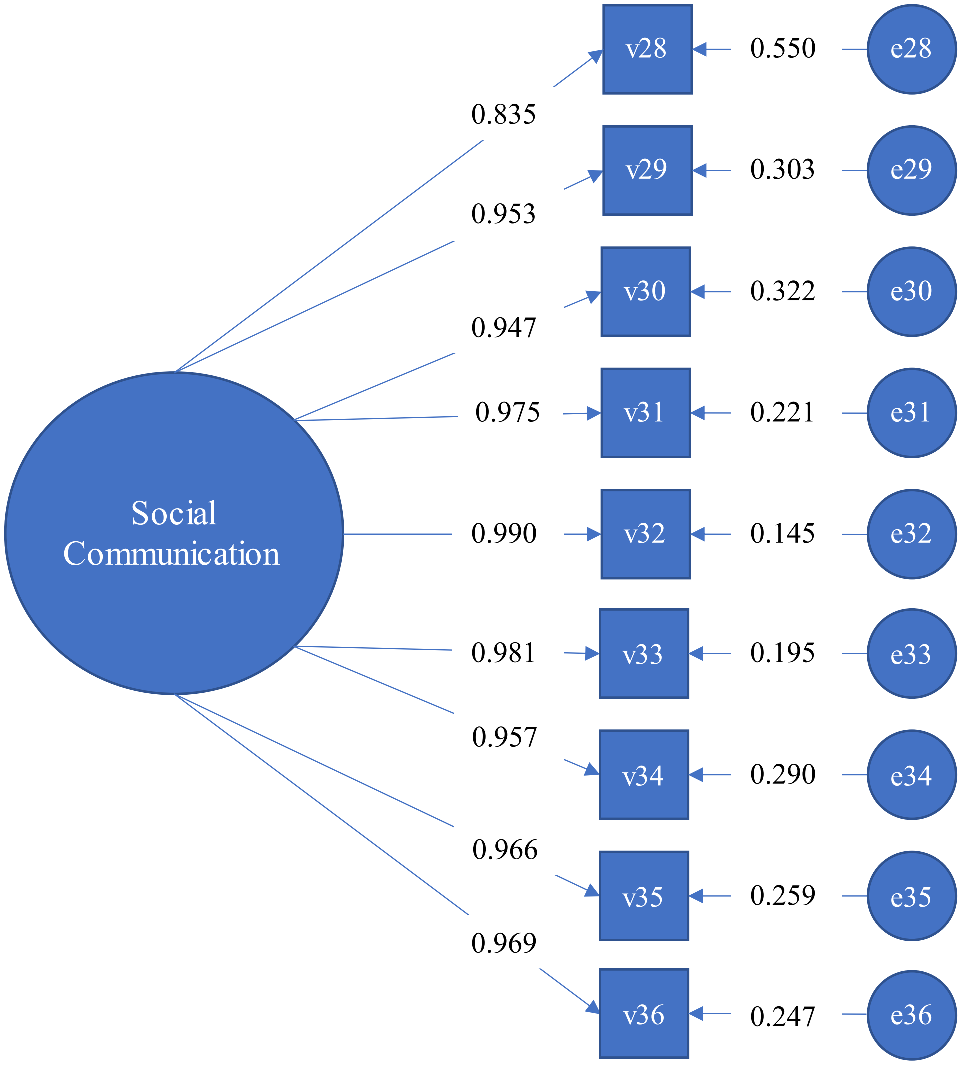


Figure D4. Path Diagram for Published GARS-3 Emotional Responses Factor

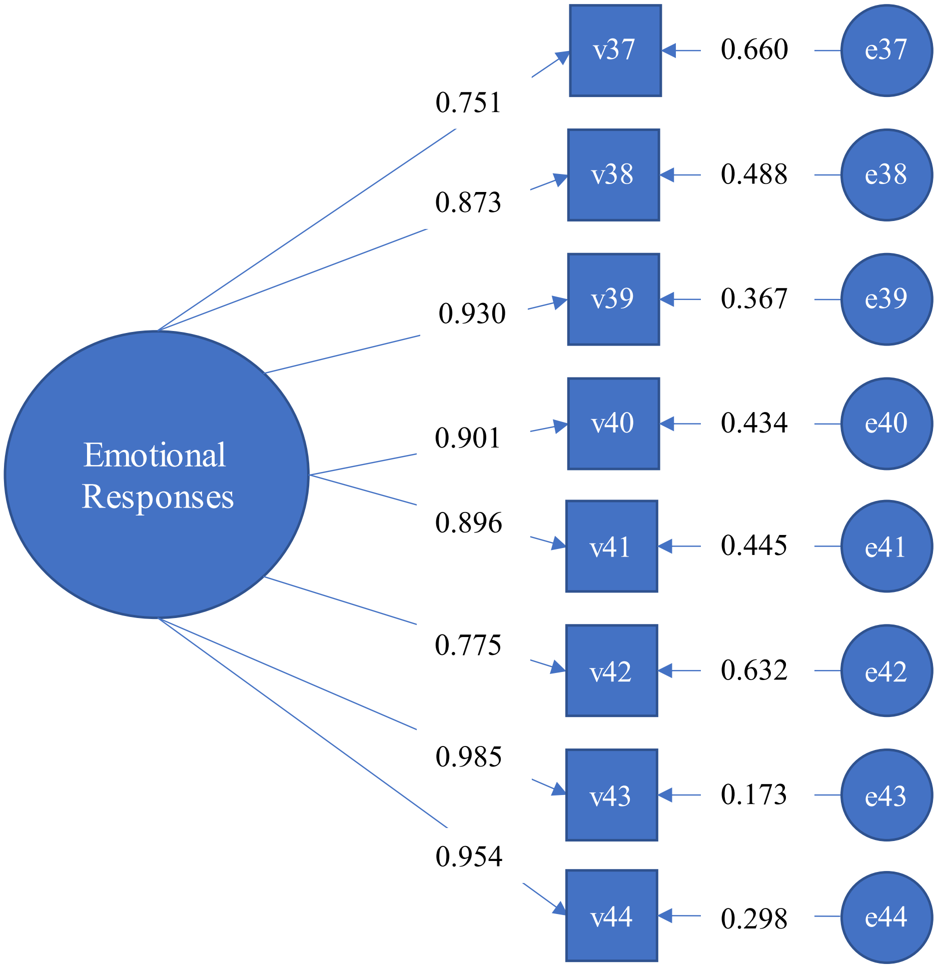


Figure D5. Path Diagram for Published GARS-3 Cognitive Style Factor

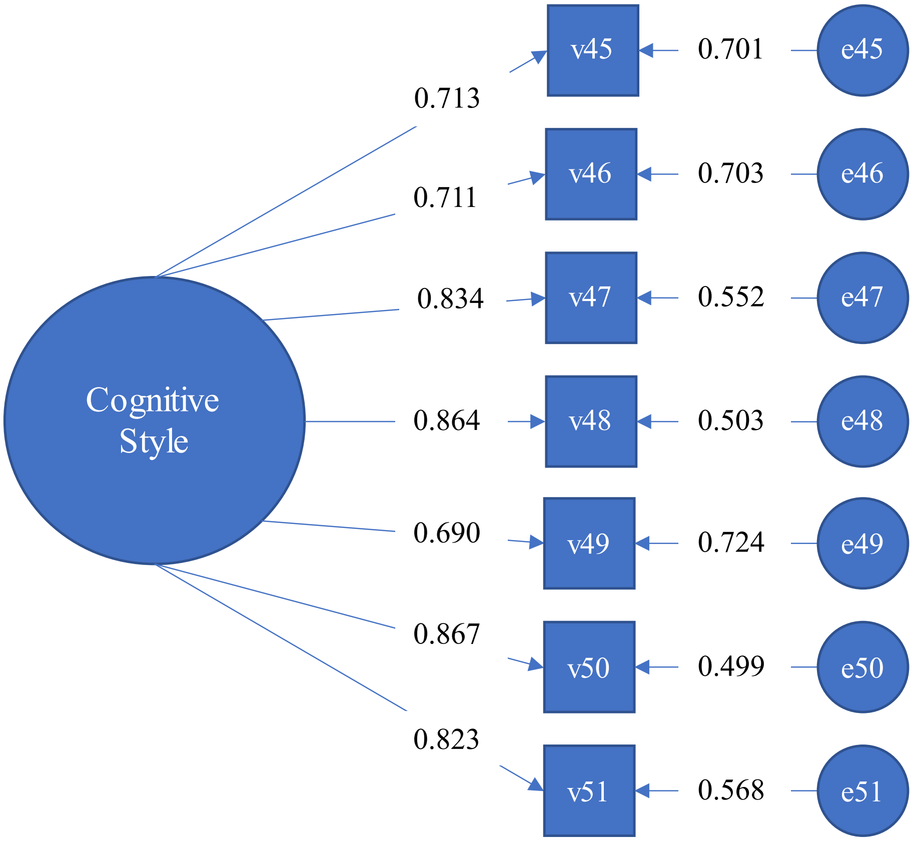
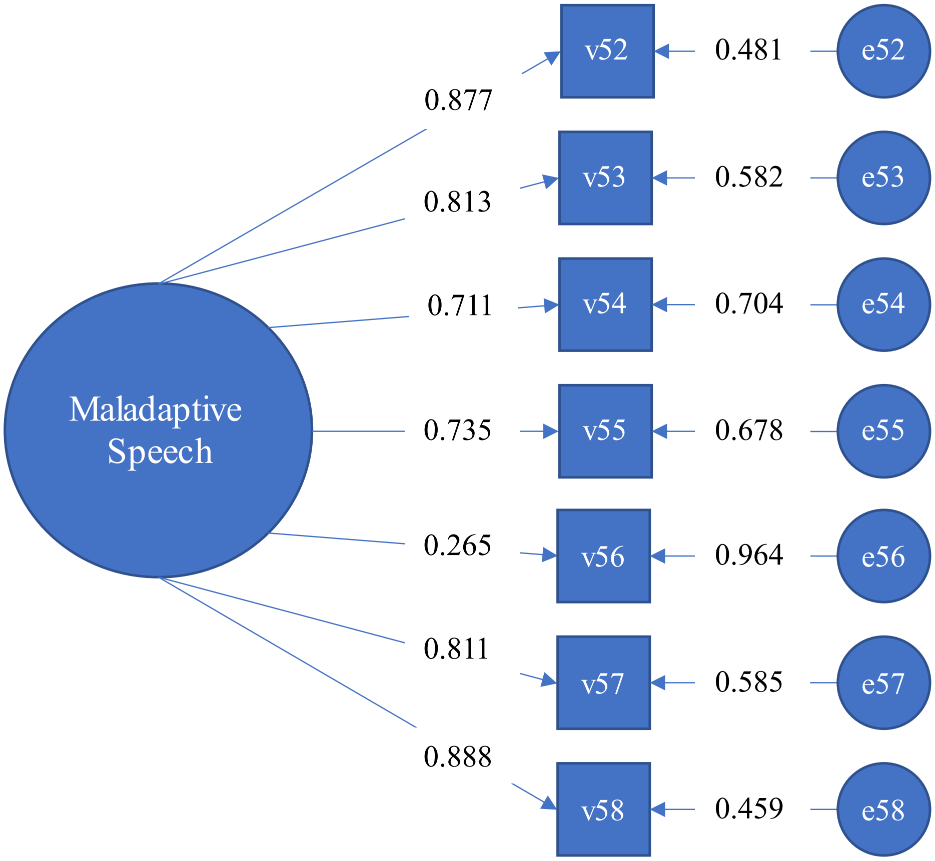


Figure D6. Path Diagram for Published GARS-3 Maladaptive Speech Factor



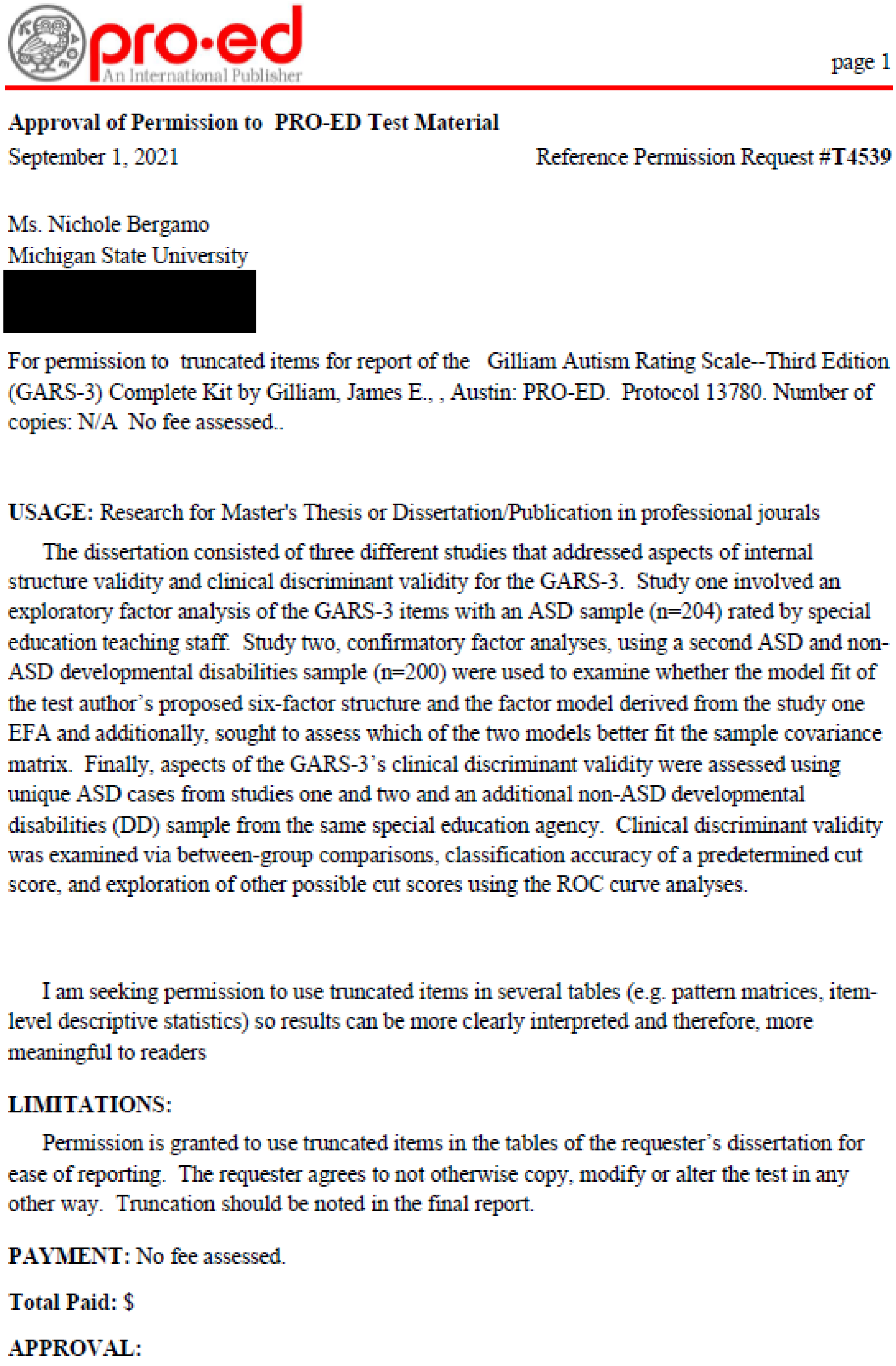
APPENDIX E: Study Two CFA Inter-Factor Correlation Matrix: Study One

Six-Factor Model

Factor Number and Name	Factor Number					
	1	2	3	4	5	6
Factor 1: Social-Emotional Reciprocity	1.00					
Factor 2: Restricted & Repetitive Behaviors	**.71**	1.00				
Factor 3: Emotion Regulation	**.50**	**.61**	1.00			
Factor 4: Social Understanding	**.84**	**.72**	**.53**	1.00		
Factor 5: Cognitive Disposition	-.12	.06	.12	-.02	1.00	
Factor 6: Speech & Language	**.50**	**.62**	**.47**	**.64**	.26	1.00

Note. Correlations greater than or equal to .30 are bolded.

APPENDIX F: Approval of Permission from PRO-ED to Report Truncated GARS-3 Items



pro·ed
An International Publisher

page 1

Approval of Permission to PRO-ED Test Material

September 1, 2021 Reference Permission Request #**T4539**

Ms. Nichole Bergamo
Michigan State University
[redacted]

For permission to truncated items for report of the   Gilliam Autism Rating Scale--Third Edition (GARS-3) Complete Kit by Gilliam, James E., , Austin: PRO-ED. Protocol 13780. Number of copies: N/A No fee assessed..

USAGE: Research for Master's Thesis or Dissertation/Publication in professional jourals

The dissertation consisted of three different studies that addressed aspects of internal structure validity and clinical discriminant validity for the GARS-3. Study one involved an exploratory factor analysis of the GARS-3 items with an ASD sample (n=204) rated by special education teaching staff. Study two, confirmatory factor analyses, using a second ASD and non-ASD developmental disabilities sample (n=200) were used to examine whether the model fit of the test author's proposed six-factor structure and the factor model derived from the study one EFA and additionally, sought to assess which of the two models better fit the sample covariance matrix. Finally, aspects of the GARS-3's clinical discriminant validity were assessed using unique ASD cases from studies one and two and an additional non-ASD developmental disabilities (DD) sample from the same special education agency. Clinical discriminant validity was examined via between-group comparisons, classification accuracy of a predetermined cut score, and exploration of other possible cut scores using the ROC curve analyses.

I am seeking permission to use truncated items in several tables (e.g. pattern matrices, item-level descriptive statistics) so results can be more clearly interpreted and therefore, more meaningful to readers

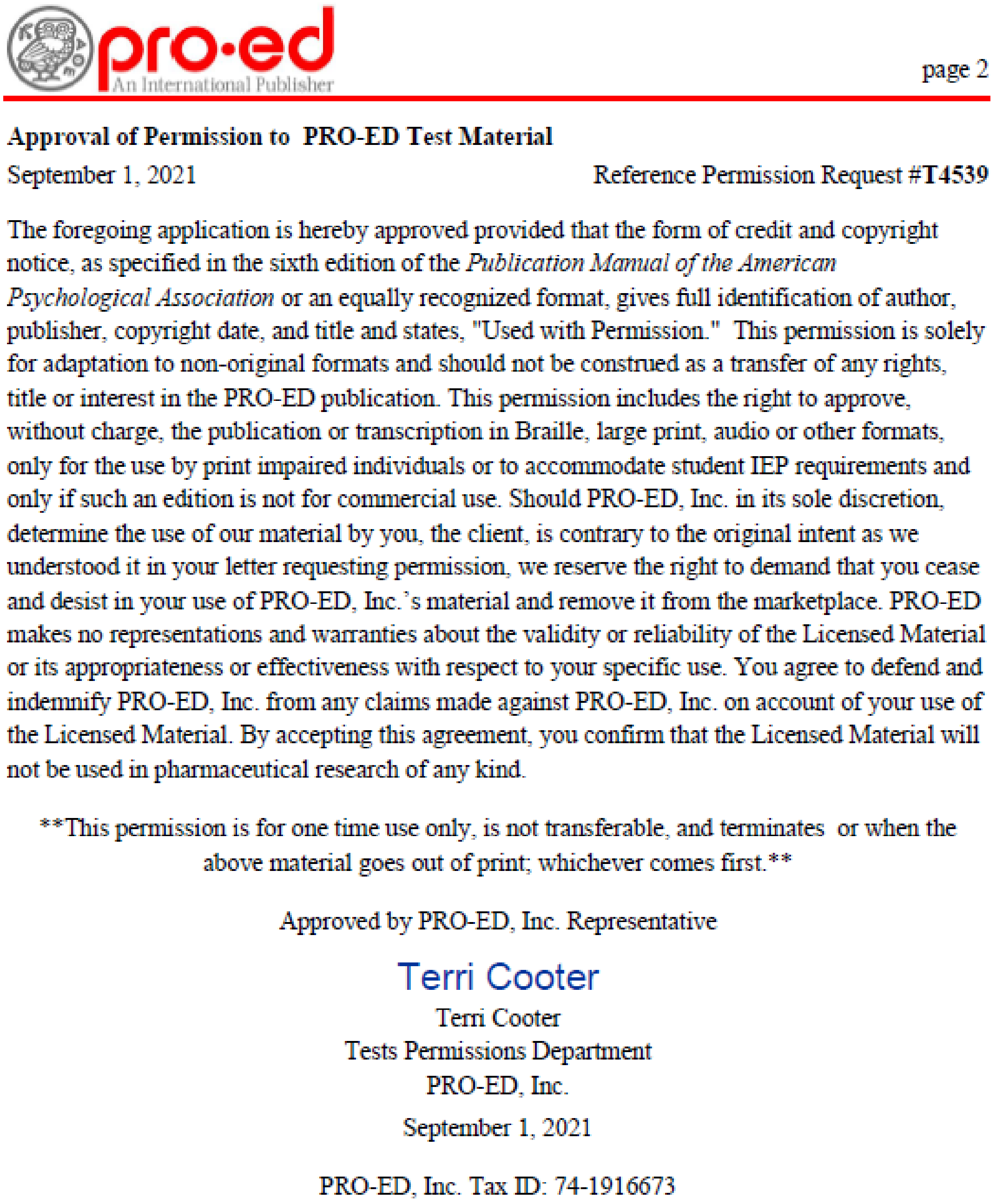
LIMITATIONS:

Permission is granted to use truncated items in the tables of the requester's dissertation for ease of reporting. The requester agrees to not otherwise copy, modify or alter the test in any other way. Truncation should be noted in the final report.

PAYMENT: No fee assessed.

Total Paid: $

APPROVAL:



pro·ed
An International Publisher

page 2

Approval of Permission to PRO-ED Test Material

September 1, 2021 Reference Permission Request #**T4539**

The foregoing application is hereby approved provided that the form of credit and copyright notice, as specified in the sixth edition of the *Publication Manual of the American Psychological Association* or an equally recognized format, gives full identification of author, publisher, copyright date, and title and states, "Used with Permission." This permission is solely for adaptation to non-original formats and should not be construed as a transfer of any rights, title or interest in the PRO-ED publication. This permission includes the right to approve, without charge, the publication or transcription in Braille, large print, audio or other formats, only for the use by print impaired individuals or to accommodate student IEP requirements and only if such an edition is not for commercial use. Should PRO-ED, Inc. in its sole discretion, determine the use of our material by you, the client, is contrary to the original intent as we understood it in your letter requesting permission, we reserve the right to demand that you cease and desist in your use of PRO-ED, Inc.'s material and remove it from the marketplace. PRO-ED makes no representations and warranties about the validity or reliability of the Licensed Material or its appropriateness or effectiveness with respect to your specific use. You agree to defend and indemnify PRO-ED, Inc. from any claims made against PRO-ED, Inc. on account of your use of the Licensed Material. By accepting this agreement, you confirm that the Licensed Material will not be used in pharmaceutical research of any kind.

This permission is for one time use only, is not transferable, and terminates or when the above material goes out of print; whichever comes first.

Approved by PRO-ED, Inc. Representative

Terri Cooter

Terri Cooter
Tests Permissions Department
PRO-ED, Inc.
September 1, 2021

PRO-ED, Inc. Tax ID: 74-1916673

REFERENCES

REFERENCES

Aiello, R., Ruble, L., & Esler, A. (2017). National study of school psychologists' use of evidence-based assessment in autism spectrum disorder. *Journal of Applied School Psychology*, *33*(1), 67-88. https://doi.org/10.1080/15377903.2016.1236307

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*, 317–322. http://cda.psych.uiuc.edu/psychometrika_highly_cited_articles/akaike_1987.pdf

Alsaedi, R. H., Carrington, S., & Watters, J. J. (2020). Behavioral and neuropsychological evaluation of executive functions in children with autism spectrum disorder in the gulf region. *Brain Sciences, 10,* 120. https://doi.org/10.3390/brainsci10020120

American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME]. (2015). *Standards for educational and psychological testing*. American Educational Research Association.

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

American Psychiatric Association. (1968). *Diagnostic and statistical manual of mental disorders* (2nd ed.). Author.

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

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed., rev.). Author.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Author.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Author.

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

Antshel, K. M., & Russo, N. (2019). Autism spectrum disorders and ADHD: Overlapping phenomenology, diagnostic issues, and treatment considerations. *Current Psychiatry Reports*, *21*, 34. https://doi.org/10.1007/s11920-019-1020-5

Ashburner, J., Ziviani, J., & Rodger, S. (2008). Sensory processing and classroom emotional, behavioral, and educational outcomes in children with autism spectrum disorder.

American Journal of Occupational Therapy, *62*, 564–573. https://doi.org/10.5014/ajot.62.5.564

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. Guilford Press.

Barnard-Brak, L., Brewer, A., Chesnut, S., Richman, D., & Schaeffer, A. M. (2016). The sensitivity and specificity of the social communication questionnaire for autism spectrum with respect to age. *Autism Research, 9,* 838–845. https://doi.org/10.1002/aur.1584

Baron-Cohen, S., Allen, J., & Gillberg, C. (1992). Can autism be detected at 18 months? The needle, the haystack, and the CHAT. *The British Journal of Psychiatry*, *161*, 839–843. https://doi.org/10.1192/bjp.161.6.839

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Statistical Psychology*, *3,* 77–85. https://doi.org/10.1111/j.2044-8317.1950.tb00285.x

Basto, M., & Pereira, J. M., (2012). An SPSS R-menu for ordinal factor analysis. *Journal of Statistical Software, 46*, 1-29.

Beck, A. T., Epstein, N., Brown, G., & Steer, R. A. (1988). An inventory for measuring clinical anxiety: Psychometric properties. *Journal of Consulting and Clinical Psychology*, *56*, 893–897. https://doi.org/10.1037/0022-006X.56.6.893

Benson, N. F., Floyd, R. G., Kranzler, J. H., Eckert, T. L., Fefer, S. A., & Morgan, G. B. (2019). Test use and assessment practices of school psychologists in the United States: Findings from the 2017 National Survey. *Journal of School Psychology*, *72*, 29–48. https://doi.org/10.1016/j.jsp.2018.12.004

Bentler, P. M. (1990). Comparative fit indices in structural models. *Psychological Bulletin*, *107*, 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Benuto, L., Zimmermann, M., Gonzalez, F., & Rodríguez, A. (2020). A confirmatory factor analysis of the beck anxiety inventory in Latinx primary care patients. *International Journal of Mental Health*, *49,* 1–21. https://doi.org/10.1080/00207411.2020.1812833

Berument, S. K., Rutter, M., Lord, C., Pickles, A., & Bailey, A. (1999). Autism screening questionnaire: Diagnostic validity. *British Journal of Psychiatry*, *175*, 444–451. https://doi.org/10.1192/bjp.175.5.444

Birnbaum, R. (2020). Exploratory and confirmatory factor analysis of the aberrant behavior checklist-community in an autism spectrum disorder sample with ratings completed by special education staff [ProQuest Information & Learning (US)]. In *Dissertation Abstracts International Section A: Humanities and Social Sciences* (Vol. 81, Issues 3-A). http://search.proquest.com/docview/2406657648/EB4FB0A94A6C467DPQ/2

Bishop, S. L., Havdahl, K. A., Huerta, M., & Lord, C. (2016). Sub-dimensions of social-communication impairment in autism spectrum disorder. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *57*, 909–916. https://doi.org/10.1111/jcpp.12510

Boothe, A., & Zuna, N. (2019). Epilepsy in children with ASD: An overview of evaluation procedures, child characteristics and treatment options. *International Journal of Disability, Development and Education*, *66*(1), 1–18. https://doi.org/10.1080/1034912X.2018.1437893

Boucher, J. (2012). Research review: Structural language in autistic spectrum disorder – characteristics and causes. *Journal of Child Psychology and Psychiatry*, *53*, 219–233. https://doi.org/10.1111/j.1469-7610.2011.02508.x

Braeken, J., & van Assen, M. A. L. M. (2017). An empirical Kaiser criterion. *Psychological Methods*, *22*, 450-466. http://doi.org/10.1037/met0000074

Breeman, S. L., Vladescu, J. C., DeBar, R. M., Grow, L. L., & Marano, K. E. (2020). The effects of procedural integrity errors during auditory–visual conditional discrimination training: A preliminary investigation. *Behavioral Interventions*, *35*, 203–216. http://doi.org/10.1002/bin.1710

Bremer, E., Graham, J. D., Heisz, J. J., & Cairney, J. (2020). Effect of acute exercise on prefrontal oxygenation and inhibitory control among male children with autism spectrum disorder: An exploratory study. *Frontiers in Behavioral Neuroscience, 14*. http://doi.org/10.3389/fnbeh.2020.00084

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). (D. A. Kenny, & T. D. Little, Eds.). Guilford Press.

Browne, M. W., & Cudeck, R. (1993). Alternate ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Sage.

Bruni, T. P. (2014). Test review: Social Responsiveness Scale – Second Edition (SRS-2). *Journal of Psychoeducational Assessment*, *32*, 365–369. https://doi.org/10.1177/0734282913517525

Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. Taylor and Francis. Kindle Edition.

Cai, R. Y., Richdale, A. L., Uljarević, M., Dissanayake, C., & Samson, A. C. (2018). Emotion regulation in autism spectrum disorder: Where we are and where we need to go. *Autism Research: Official Journal of the International Society for Autism Research*, *11*, 962–978. https://doi.org/10.1002/aur.1968

Campanaro, A. M., Vladescu, J. C., Kodak, T., DeBar, R. M., & Nippes, K. C. (2020). Comparing skill acquisition under varying onsets of differential reinforcement: A preliminary analysis. *Journal of Applied Behavior Analysis*, *53*, 690–706. http://doi.org/10.1002/jaba.615

Carlile, K. A., DeBar, R. M., Reeve, S. A., Reeve, K. F., & Meyer, L. S. (2018). Teaching help-seeking when lost to individuals with autism spectrum disorder. *Journal of Applied Behavior Analysis, 51*, 191–206. http://doi.org/10.1002/jaba.447

Cattell, R. B. (1966). The Scree Test for the Number of Factors. *Multivariate Behavioral Research*, *1*, 245–276. https://doi.org/10.1207/s15327906mbr0102_10

Centers for Disease Control and Prevention (CDC). (2020). Autism data visualization tool. https://www.cdc.gov/ncbddd/autism/data/index.html.

Centers for Disease Control and Prevention (CDC). (2019). Signs and symptoms of autism spectrum disorder. https://www.cdc.gov/ncbddd/autism/signs.html

Centers for Disease Control and Prevention (CDC). (2014). Prevalence of autism spectrum disorder among children aged 8 years—Autism and Developmental Disabilities Monitoring Network, 11 sites, United States, 2010, *MMWR Surveillance Summaries, 63,* (No. SS-2). http://doi.org/10.15585/mmwr.ss6706a1

Centers for Disease Control and Prevention (CDC). (2007). Prevalence of autism spectrum disorders – Autism and Developmental Disabilities Monitoring Network, six sites, United States, 2000, *MMWR Surveillance Summaries*, *56*, 1-11. https://www.cdc.gov/mmwr/preview/mmwrhtml/ss5601a1.htm

Chandler, S., Charman, T., Baird, G., Simonoff, E., Loucas, T., Meldrum, D. et al. (2007) Validation of the Social Communication Questionnaire in a population cohort of children with autism spectrum disorders. Journal of the American Academy of Child and Adolescent Psychiatry, *46*, 1324–1332. https://doi.org/10.1097/chi.0b013e31812f7d8d

Clark, E., Radley, K. C., & Phosaly, L. (2014). Best practices in assessment and intervention of children with high-functioning autism spectrum disorders. *Best practices in school psychology: Data-based and collaborative decision making*, 417-431.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Earlbaum Associates.

Constantino, J. N., & Gruber, C. P. (2012). *Social Responsiveness Scale, Second Edition (SRS-2).* Western Psychological Services.

Corsello, C., Hus, V., Pickles, A., Risi, S., Cook Jr, E. H., Leventhal, B. L., & Lord, C. (2007). Between a ROC and a hard place: decision making and making decisions about using the SCQ. *Journal of Child Psychology and Psychiatry, 48,* 932-940. https://doi.org/10.1111/j.1469-7610.2007.01762.x

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation, 10,* 1-9. https://doi.org/10.7275/JYJ1-4868

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334. https://doi.org/10.1007/BF02310555

Cubicciotti, J. E., Vladescu, J. C., Reeve, K. F., Carroll, R. A., & Schnell, L. K. (2019). Effects of stimulus presentation order during auditory–visual conditional discrimination training for children with autism spectrum disorder. *Journal of Applied Behavior Analysis, 52,* 541–556. http://doi.org/10.1002/jaba.530

Dass, T. K., Kisamore, A. N., Vladescu, J. C., Reeve, K. F., Reeve, S. A., & Taylor-Santa, C. (2018). Teaching children with autism spectrum disorder to tact olfactory stimuli. *Journal of Applied Behavior Analysis*, *51,* 538–552. http://doi.org/10.1002/jaba.470

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the *EM* algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

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,* 425-438. https://doi.org/10.1080/10705511.2014.915373

Duffy, L., Baluch, B., Welland, S., & Raman, E. (2017). Effects of physical activity on debilitating behaviours in 13- to 20-year-old males with severe autism spectrum disorder. *Journal of Exercise Rehabilitation*, *13,* 340–347. https://doi.org/10.12965/jer.1734960.480

Dunn, D. M. (2018). *Peabody Picture Vocabulary Test* (5th ed.). Pearson.

Elliott, C. D. (2007). *Differential ability scales* (2nd ed.). The Psychological Corporation.

Evans, B. (2013). How autism became autism: The radical transformation of a central concept of child development in Britain. *History of the Human Sciences*, *26,* 3–31. https://doi.org/10.1177/0952695113484320

Eskow, K. G., Link to external site, this link will open in a new window, Chasson, G. S., & Summers, J. A. (2019). The role of choice and control in the impact of autism waiver services on family quality of life and child progress. *Journal of Autism and Developmental Disorders; New York, 49*, 2035–2048. http://doi.org/10.1007/s10803-019-03886-5

Ezzeddine, E. W., DeBar, R. M., Reeve, S. A., & Townsend, D. B. (2020). Using video modeling to teach play comments to dyads with ASD. *Journal of Applied Behavior Analysis*, *53*, 767–781. https://doi.org/10.1002/jaba.621

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

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

Fombonne, E. (2002). Epidemiological trends in rates of autism. *Molecular Psychiatry*, *7*, S4–S6. https://doi.org/10.1038/sj.mp.4001162

Fombonne E. (2003). The prevalence of autism. *Journal of AMA*, *289*(1), 87-89. https://doi.org/10.1001/jama.289.1.87

Frazier, T. W., Youngstrom, E. A., Speer, L., Embacher, R., Law, P., Constantino, J., Findling, R. L., Hardan, A. Y., & Eng, C. (2012). Validation of proposed DSM-5 criteria for autism spectrum disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, *51*(1), 28-40.e3. https://doi.org/10.1016/j.jaac.2011.09.021

Gaskin, C. J., Lambert, S. D., Bowe, S. J., & Orellana, L. (2017). Why sample selection matters in exploratory factor analysis: Implications for the 12-item World Health Organization Disability Assessment Schedule 2.0. *BMC Medical Research Methodology, 17*. https://doi.org/10.1186/s12874-017-0309-5

Gerbing, D. W., & Hamilton, J. G. (1996). Viability of exploratory factor analysis as a precursor to confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *3*(1), 62–72. https://doi.org/10.1080/10705519609540030

Gilliam, J. E. (1995). *Gilliam Autism Rating Scale (GARS)*. Pro-Ed.

Gilliam, J. E. (2006). *Gilliam Autism Rating Scale-Second Edition (GARS-2)*. Pro-Ed.

Gilliam, J. E. (2013). *Gilliam Autism Rating Scale–Third Edition (GARS-3)*. Pro-Ed.

Goldstein, S., & Naglieri, J. A. (2009). *Autism Spectrum Rating Scales (ASRS)*. Multi-Health Systems.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Lawrence Erlbaum Associates, Inc.

Guy, W. (1976). *ECDEU Assessment Manual for Psychopharmacology*. U.S. Department of Health, Education, and Human Welfare.

Harker, C. M., & Stone, W. L. (2014). Comparison of the diagnostic criteria for autism spectrum disorder across DSM-5, DSM-IV-TR, and the Individuals with Disabilities Education Act

(IDEA) definition of autism. *The Iris Center*. https://iris.peabody.vanderbilt.edu/wp-content/uploads/pdf_info_briefs/ASD_Comparison_information_brief.pdf

Hastings, K., & Campbell, J. (2016, May 13). *An initial evaluation of the validity of the Gilliam Autism Rating Scale-Third Edition (GARS-3) in a clinical sample*. Poster presentation.

Hoff, K. E. & Doepke, K, J, (2014). Review of the Social Responsiveness Scale, Second Edition. In Carlson, J. F., Geisinger, K. F., & Jonson, J. L. (Eds.), *The Nineteenth Mental Measurements Yearbook* (pp. 637-639). Buros Center for Testing.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179–185. https://doi.org/10.1007/BF02289447

Hu, L., & 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. https://doi.org/10.1080/10705519909540118

Huerta, M., & Lord, C. (2012). Diagnostic evaluation of autism spectrum disorders. *Pediatric Clinics of North America*, *59*(1), 103–111. https://doi.org/10.1016/j.pcl.2011.10.018

Hutchins, T. L. (2017). Review of the Gilliam Autism Rating Scale – Third Edition. In J. F. Carlson, K. F. Geisinger, & J. L. Jonson (Eds.), *The Twentieth Mental Measurements Yearbook* (pp. 376-378). Buros Center for Testing.

Hyman, S. L., Levy, S. E., & Myers, S. M. (2020). Executive summary: Identification, evaluation, and management of children with autism spectrum disorder. *Pediatrics, 145*(1), e20193448. https://doi.org/10.1542/peds.2019-3448

IBM Corp. (2019a). *IBM SPSS Missing Values 26*. IBM Corp.

IBM Corp. (2019b). *IBM SPSS Statistics for Windows, Version 26.0*. IBM Corp.

IBM Corp. (2010). *SPSS R Plug-in 2.10*. IBM Corp.

Jang, J., Matson, J. L., Williams, L. W., Tureck, K., Goldin, R. L., & Cervantes, P. E. (2013). Rates of comorbid symptoms in children with ASD, ADHD, and comorbid ASD and ADHD. *Research in Developmental Disabilities*, *34*, 2369–2378. http://doi.org./10.1016/j.ridd.2013.04.021

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika, 35*, 401-415. https://doi.org/10.1007/BF02291817

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20,* 141–151. https://doi.org/10.1177/001316446002000116

Kaiser, H. F., & Rice, J. (1974) Little jiffy, mark iv. *Educational and Psychological Measurement, 34*(1), 111-117. https://doi.org/10.1177/001316447403400115

Kamp-Becker, I., Tauscher, J., Wolff, N., Küpper, C., Poustka, L., Roepke, S., Roessner, V., Heider, D., & Stroth, S. (2021). Is the combination of ADOS and ADI-R necessary to classify ASD? Rethinking the "gold standard" in diagnosing ASD. *Frontiers in Psychiatry, 12*. https://www.frontiersin.org/article/10.3389/fpsyt.2021.727308

Kanner, L. (1943). Autistic disturbances of affective contact. *Nervous Child, 2,* 217–250.

Karren, B. C. (2017). A test review: Gilliam, J. E. (2014). Gilliam Autism Rating Scale–Third Edition (GARS-3). *Journal of Psychoeducational Assessment, 35,* 342–346. https://doi.org/10.1177/0734282916635465

Kay, J. C., Kisamore, A. N., Vladescu, J. C., Sidener, T. M., Reeve, K. F., Taylor-Santa, C., & Pantano, N. A. (2020). Effects of exposure to prompts on the acquisition of intraverbals in children with autism spectrum disorder. *Journal of Applied Behavior Analysis, 53*(1), 493–507. http://doi.org/10.1002/jaba.606

Kazdin, A. E. (2017). *Research design in clinical psychology* (5th ed.). Pearson.

Kidd, S. A., Berry-Kravis, E., Choo, T. H., Chen, C., Esler, A., Hoffmann, A., Andrews, H. F., & Kaufmann, W. E. (2020). Improving the diagnosis of autism spectrum disorder in Fragile X syndrome by adapting the Social Communication Questionnaire and the Social Responsiveness Scale-2. *Journal of Autism and Developmental Disorders, 50,* 3276–3295. https://doi.org/10.1007/s10803-019-04148-0

Kim, M., Winkler, C., & Talley, S. (2021). Binary item CFA of Behavior Problem Index (BPI) using Mplus: A step-by-step tutorial. *The Quantitative Methods for Psychology, 17,* 141–153. https://doi.org/10.20982/tqmp.17.2.p141

Kluck, A. S. (2014). Review of the Autism Spectrum Rating Scales. In Carlson, J. F., Geisinger, K. F., & Jonson, J. L. (Eds.), *The Nineteenth Mental Measurements Yearbook* (pp. 32-34). Buros Center for Testing.

Knowland, V. C. P., Fletcher, F., Henderson, L.-M., Walker, S., Norbury, C. F., & Gaskell, M. G. (2019). Sleep promotes phonological learning in children across language and autism spectra. *Journal of Speech, Language, and Hearing Research, 62,* 4235–4255. http://doi.org/10.1044/2019_JSLHR-S-19-0098

Krug, D. A., & Arick, J. R. (2003). *KADI: Krug Asperger's disorder index*. PRO-ED.

Kuriakose, S., & Shalev, R. (2016). Early diagnostic assessment. In R. Lang, T. B. Hancock, & N. N. Singh (Eds.), *Early Intervention for Young Children with Autism Spectrum Disorder* (pp. 15–46). Springer International Publishing. https://doi.org/10.1007/978-3-319-30925-5_2

Lai, M.-C., Lombardo, M. V., Auyeung, B., Chakrabarti, B., & Baron-Cohen, S. (2015). Sex/gender differences and autism: Setting the scene for future research. *Journal of the American Academy of Child & Adolescent Psychiatry*, *54*(1), 11–24. https://doi.org/10.1016/j.jaac.2014.10.003

Lalkhen, A. G., & McCluskey, A. (2008). Clinical tests: Sensitivity and specificity. *Continuing Education in Anaesthesia Critical Care & Pain*, *8*, 221–223. https://doi.org/10.1093/bjaceaccp/mkn041

Lecavalier, L. (2005). An evaluation of the Gilliam autism rating scale. *Journal of Autism and Developmental Disorders, 35,* 795-813. https://doi.org/10.1007/s10803-005-0025-6

Lieneman, C. C., Ruckle, M. M., & McNeil, C. B. (2018). Parent-child interaction therapy for a child with autism spectrum disorder: A case study examining effects on ASD symptoms, social engagement, pretend play, and disruptive behavior. In C. B. McNeil, L. B. Quetsch, & C. M. Anderson (Eds.), *Handbook of parent-child interaction therapy for children on the autism spectrum* (pp. 677–696). Springer International Publishing. https://doi.org/10.1007/978-3-030-03213-5_39

Lord, C., Rutter, M., DiLavore, P., & Risi, S. (1999). *Autism Diagnostic Observation Schedule (ADOS)*. Western Psychological Services.

Lord, C., Rutter, M., DiLavore, P. C., Risi, S., Gotham, K., Bishop, S. (2012). *Autism Diagnostic Observation Schedule, Second Edition (ADOS-2)*. Western Psychological Services.

Lord, C., Rutter, M., & LeCouteur, 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,* 659–685. https://doi.org/10.1007/BF02172145

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

Lord, C., Wagner, A., Rogers, S., Szatmari, P., Aman, M., Charman, T., Dawson, G., Durand, V. M., Grossman, L., Guthrie, D., Harris, S., Kasari, C., Marcus, L., Murphy, S., Odom, S., Pickles, A., Scahill, L., Shaw, E., Siegel, B., … Yoder, P. (2005). Challenges in evaluating psychosocial interventions for autistic spectrum disorders. *Journal of Autism and Developmental Disorders*, *35*, 695–708. https://doi.org/10.1007/s10803-005-0017-6

Lordo, D., Bertolin, M., Sudikoff, E., Keith, C., Braddock, B., & Kaufman, D. A. S. (2017). Parents perceive improvements in socio-emotional functioning in adolescents with ASD following social skills treatment. *Journal of Autism and Developmental Disorders*, *47*(1), 203–214. http://doi.org/10.1007/s10803-016-2969-0

Loucas, T., Charman, T., Pickles, A., Simonoff, E., Chandler, S., Meldrum, D., & Baird, G. (2008). Autistic symptomatology and language ability in autism spectrum disorder and specific language impairment. *Journal of Child Psychology and Psychiatry, 49*, 1184–1192. https://doi.org/10.1111/j.1469-7610.2008.01951.x

Maenner, M. J., Schieve, L. A., Rice, C. E., Cunniff, C., Giarelli, E., Kirby, R. S., Lee, L.-C., Nicholas, J. S., Wingate, M. S., & Durkin, M. S. (2013). Frequency and pattern of documented diagnostic features and the age of autism identification. *Journal of the American Academy of Child & Adolescent Psychiatry*, *52*, 401-413.e8. https://doi.org/10.1016/j.jaac.2013.01.014

Maenner, M.J., Shaw, K.A., Bakian, A.V., et al. (2021). Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2018. *MMWR Surveillance Summaries*, 70 (No. SS-11), 1–16. https://doi.org/10.15585/mmwr.ss7011a1

Maenner M.J., Shaw, K.A., Baio J., et al. (2020). Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and Developmental Disabilities Monitoring Network, 11 sites, United States, 2016. *MMWR Surveillance Summaries, 69* (No. SS-4), 1–12. https://doi.org/10.15585/mmwr.ss6904a1

Magyar, C. I., Pandolfi, V., & Dill, C. A. (2012). An initial evaluation of the Social Communication Questionnaire for the assessment of autism spectrum disorders in children with Down syndrome. *Journal of Developmental & Behavioral Pediatrics*, *33*, 134–145. https://doi.org/10.1097/DBP.0b013e318240d3d9

Malcolm, K. K. (2014). Review of the Childhood Autism Rating Scale, Second Edition. In Carlson, J. F., Geisinger, K. F., & Jonson, J. L. (Eds.), *The Nineteenth Mental Measurements Yearbook* (pp. 143-146). Buros Center for Testing.

Mandell, D. S., Lawer, L. J., Branch, K., Brodkin, E. S., Healey, K., Witalec, R., Johnson, D. N., & Gur, R. E. (2012). Prevalence and correlates of autism in a state psychiatric hospital. *Autism*, *16*, 557–567. https://doi.org/10.1177/1362361311412058

March, J. S., Conners, C., Arnold, G., Epstein, J., Parker, J., Hinshaw, S., Abikoff, H., Molina, B., Wells, K., Newcorn, J., Schuck, S., Pelham, W. E., & Hoza, B. (1999). The Multidimensional Anxiety Scale for Children (MASC): Confirmatory factor analysis in a pediatric ADHD sample. *Journal of Attention Disorders, 3*, 85–89. https://doi.org/10.1177/108705479900300202

Martin, N., & Brownell, R. (2010). *Expressive One-Word Picture Vocabulary Test* (4th ed.). Academic Therapy Publications.

Matheis, M., Matson, J. L., Hong, E., & Cervantes, P. E. (2019). Gender differences and similarities: Autism symptomatology and developmental functioning in young children.

Journal of Autism and Developmental Disorders, *49*, 1219–1231. https://doi.org/10.1007/s10803-018-3819-z

Matson, J. L., & Cervantes, P. E. (2014). Commonly studied comorbid psychopathologies among persons with autism spectrum disorder. *Research in Developmental Disabilities, 35,* 952–962. https://doi.org/10.1016/j.ridd.2014.02.012

Matson, J. L., & Shoemaker, M. (2009). Intellectual disability and its relationship to autism spectrum disorders. *Research in Developmental Disabilities, 30,* 1107–1114. https://doi.org/10.1016/j.ridd.2009.06.003

MacCallum, R. C., Widaman, K., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*(1), 84-99. https://doi.org/10.1037/1082-989X.4.1.84

McClellan, M. J. (2014). Review of the Childhood Autism Rating Scale, Second Edition. In Carlson, J. F., Geisinger, K. F., & Jonson, J. L. (Eds.), *The Nineteenth Mental Measurements Yearbook* (pp. 146-147). Buros Center for Testing.

McCrimmon, A., & Rostad, K. (2014). Test review: Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) manual (part II): Toddler module. *Journal of Psychoeducational Assessment*, *32*(1), 88–92. https://doi.org/10.1177/0734282913490916

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine, 8,* 283–298. https://doi.org/10.1016/S0001-2998(78)80014-2

Minaei, A., & Nazeri, S. (2018). Psychometric properties of the Gilliam Autism Rating Scale–Third Edition (GARS-3) in individuals with autism: A pilot study. *Journal of Exceptional Children*, *18*, 113–122. http://joec.ir/article-1-847-en.html

Montgomery, J. M., Newton, B., & Smith, C. (2008). Test Review: Gilliam, J. (2006). GARS-2: Gilliam Autism Rating Scale—Second Edition. Austin, TX: PRO-ED. *Journal of Psychoeducational Assessment*, *26*, 395–401. https://doi.org/10.1177/0734282908317116

Moulton, E., Bradbury, K., Barton, M., & Fein, D. (2019). Factor analysis of the childhood autism rating scale in a sample of two year olds with an autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *49*, 2733–2746. https://doi.org/10.1007/s10803-016-2936-9

Mullen, E. (1995). *The Mullen Scales of Early Learning*. American Guidance Service.

Murphy, K. R., & Davidshofer, C. O. (2005). *Psychological testing: Principles and applications* (6th ed.). Pearson Education.

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Myles, B. S., Jones-Bock, S., & Simpson, R. L. (2001). *Asperger syndrome diagnostic scale*. PRO-ED.

Nasca, B. C., Lopata, C., Donnelly, J. P., Rodgers, J. D., & Thomeer, M. L. (2019). Sex differences in externalizing and internalizing symptoms of children with ASD. *Journal of Autism and Developmental Disorders, 50,* 3245–3252. https://doi.org/10.1007/s10803-019-04132-8

Nelson, A. T., Lopata, C., Volker, M. A., Thomeer, M.L., Toomey, J. A., & Dua, E. (2016). Exploratory factor analysis of SRS-2 teacher ratings for youth with ASD. *Journal of Autism and Developmental Disorders, 46,* 2905–2915. https://doi.org/10.1007/s10803-016-2822-5

Noland, R. M., & Gabriels, R. L. (2004). Screening and identifying children with autism spectrum disorders in the public school system: The development of a model process. *Journal of Autism and Developmental Disorders*, *34*, 265-277. https://doi.org/10.1023/B:JADD.0000029549.84385.44

Northgrave, J., Vladescu, J. C., DeBar, R. M., Toussaint, K. A., & Schnell, L. K. (2019). Reinforcer choice on skill acquisition for children with autism spectrum disorder: A systematic replication. *Behavior Analysis in Practice, 12,* 401–406. http://doi.org/10.1007/s40617-018-0246-8

Norris, M., & Lecavalier, L. (2010a). Evaluating the use of exploratory factor analysis in developmental disability psychological research. *Journal of Autism and Developmental Disorders, 40*(1), 8–20. https://doi.org/10.1007/s10803-009-0816-2

Norris, M., & Lecavalier, L. (2010b). Screening accuracy of level 2 autism spectrum disorder rating scales: A review of selected instruments. *Autism*, *14*, 263–284. https://doi.org/10.1177/1362361309348071

Norris, M., Lecavalier, L., & Edwards, M. C. (2012). The structure of autism symptoms as measured by the autism diagnostic observation schedule. *Journal of Autism and Developmental Disorders, 42,* 1075–1086. http://doi.org/10.1007/s10803-011-1348-0

Nunnally, J. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

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

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers, 32,* 396–402. https://doi.org/10.3758/BF03200807

Osborne, J. W., & Banjanovic, E. S. (2016). *Exploratory factor analysis with SAS* ®. SAS Institute Inc. https://doi.org/10.1207/s15374424jccp3403_8

Pandolfi, V., & Magyar, C. I. (2016). Psychopathology. In J. L. Matson (Ed.), *Comorbid Conditions Among Children with Autism Spectrum Disorders* (pp. 171–186). Springer International Publishing. https://doi.org/10.1007/978-3-319-19183-6_7

Pandolfi, V., Magyar, C. I., & Dill, C. A. (2010). Constructs assessed by the GARS-2: factor analysis of data from the standardization sample. *Journal of Autism and Developmental Disorders, 40,* 1118-1130. https://doi.org/10.1007/s10803-010-0967-1

Peters-Scheffer, N., Didden, R., & Lang, R. (2016). Intellectual disability. In J. L. Matson (Ed.), *Comorbid Conditions Among Children with Autism Spectrum Disorders* (pp. 283–300). Springer International Publishing. https://doi.org/10.1007/978-3-319-19183-6_12

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*. Sage.

Pfeiffer, B., Erb, S. R., & Slugg, L. (2019a). Impact of noise-attenuating headphones on participation in the home, community, and school for children with autism spectrum disorder. *Physical & Occupational Therapy in Pediatrics*, *39*, 60–76. https://doi.org/10.1080/01942638.2018.1496963

Pfeiffer, B., Piller, A., Bevans, K., & Shiu, C. (2019b). Reliability of the participation and sensory environment questionnaire: Community scales. *Research in Autism Spectrum Disorders*, *64*, 84–93. https://doi.org/10.1016/j.rasd.2019.03.008

Pfeiffer, B., Piller, A., Slugg, L., & Shiu, C. (2018). Brief report: Reliability of the participation and sensory environment questionnaire: Home scales. *Journal of Autism and Developmental Disorders; New York, 48*, 2567–2576. http://doi.org/10.1007/s10803-018-3499-8

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for 267 Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/

Raykov, T., & Marcoulides, G. A. (2008). *An introduction to applied multivariate analysis*. Taylor & Francis Group.

Rispoli, M., Brodhead, M., Wolfe, K., & Gregori, E. (2018). Trial-based functional analysis informs treatment for vocal scripting. *Behavior Modification*, *42*, 441–465. https://doi.org/10.1177/0145445517742882

Robins, D. L., Casagrande, K., Barton, M., Chen, C. M. A., Dumont-Mathieu, T., & Fein, D. (2014). Validation of the modified checklist for autism in toddlers, revised with follow-up (M-CHAT-R/F). *Pediatrics, 133*(1), 37-45. https://doi.org/10.1542/peds.2013-1813

Robins, D., Fein, D., & Barton M. (2009). *Modified checklist for autism in toddlers, revised with follow-up (M-CHAT-R/F).* Self-published.

Rossi, M. R., Vladescu, J. C., Reeve, K. F., & Gross, A. C. (2017). Teaching safety responding to children with autism spectrum disorder. *Education & Treatment of Children*, *40*, 187–208. http://doi.org/10.1353/etc.2017.0009

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology*, *47*, 537–560. https://doi.org/10.1111/j.1744-6570.1994.tb01736.x

Rutter, M. (2005). Incidence of autism spectrum disorders: Changes over time and their meaning. *Acta Paediatrica*, *94*(1), 2–15. https://doi.org/10.1111/j.1651-2227.2005.tb01779.x

Rutter, M., Bailey, A., & Lord, C. (2003a). *Social Communication Questionnaire (SCQ)*. Western Psychological Services.

Rutter, M., Le Couteur, A., Lord, C. (2003b). *Autism Diagnostic Interview – Revised (ADI-R)*. Western Psychological Services.

Salvia, J., Ysseldyke, J., & Witmer, S. (2017). *Assessment in Special and Inclusive Education* (13th edition). Boston, MA: Cengage Learning.

SAS Institute, Inc. (2013). *SAS version 9.4.* SAS Institute Inc.

Sattler, J. M. (2008). *Assessment of children: Cognitive Foundations* (5th ed., p. 109). Jerome M. Sattler, Publisher, Inc.

Scahill, L., & Lord, C. (2004). Subject selection and characterization in clinical trials in children with autism. *CNS Spectrums, 9*(1), 22-32. https://www.cambridge.org/core/product/identifier/S1092852900008336/type/journal_article

Schopler, E., Van Bourgondien, M. E., Wellman, G. J., & Love, S. R. (2010). *The childhood autism rating scale, Second Edition* (CARS2). Western Psychological Services.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464. https://doi.org/10.1214/aos/1176344136

Shaw, S. R. (2014). Review of the Autism Spectrum Rating Scales. In Carlson, J. F., Geisinger, K. F., & Jonson, J. L. (Eds.), *The Nineteenth Mental Measurements Yearbook* (pp. 34-37). Buros Center for Testing.

Shea, V., & Mesibov, G. B. (2005). Adolescents and adults with autism. In F. R. Volkmar, R. Paul, A. Klin, & D. Cohen (Eds.), *Handbook of autism and pervasive developmental disorders: Vol. 1. Diagnosis development, neurobiology, and behavior* (3rd ed., pp. 288–311). John Wiley & Sons.

Simek, A. N., & Wahlberg, A. C. (2011). Test review: Autism Spectrum Rating Scales. *Journal of Psychoeducational Assessment*, *29*, 191–95. https://doi.org/10.1177/0734282910375408

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 & Adolescent Psychiatry*, *47*, 921–929. http://doi.org/10.1097/CHI.0b013e318179964f

Smith, T., Scahill, L., Dawson, G., Guthrie, D., Lord, C., Odom, S., Rogers, S., & Wagner, A. (2007). Designing research studies on psychosocial interventions in autism. *Journal of Autism and Developmental Disorders*, *37*, 354–366. https://doi.org/10.1007/s10803-006-0173-3

South, M., Williams, B. J., McMahon, W. M., Owley, T., Filipek, P. A., Shernoff, E., Corsello, C., Lainhart, J. E., Landa, R., & Ozonoff, S. (2002). Utility of the Gilliam autism rating scale in research and clinical populations. *Journal of Autism and Developmental Disorders*, *32*, 593-599. http://doi.org/10.1023/a:1021211232023

Sparrow, S. S., Cicchetti, D. V., & Saulnier, C. A. (2016). *Vineland Adaptive Behavior Scales – Third Edition*. Pearson.

Steiger, J. H., & Lind, J. M. (1980, June). *Statistically based tests for the number of common factors*. Paper presented at the meeting of the Psychometric Society, Iowa City, IA.

Stevens, T., Peng, L., & Barnard-Brak, L. (2016). The comorbidity of ADHD in children diagnosed with autism spectrum disorder. *Research in Autism Spectrum Disorders, 31*, 11–18. https://doi.org/10.1016/j.rasd.2016.07.003

Syriopoulou-Delli, C. K., Agaliotis, I., & Papaefstathiou, E. (n.d.). Social skills characteristics of students with autism spectrum disorder. *International Journal of Developmental Disabilities*, *64*, 35–44. https://doi.org/10.1080/20473869.2016.1219101

Talbott, M. R., Young, G. S., Munson, J., Estes, A., Vismara, L. A., & Rogers, S. J. (2020). The developmental sequence and relations between gesture and spoken language in toddlers with autism spectrum disorder. *Child Development*, *91*, 743–753. https://doi.org/10.1111/cdev.13203

Tager-Flusberg, H., & Kasari, C. (2013). Minimally verbal school-aged children with autism spectrum disorder: The neglected end of the spectrum. *Autism Research, 6*, 468–478. https://doi.org/10.1002/aur.1329

Tager-Flusberg, H., Paul, R., & Lord, C. (2005). Language and Communication in Autism. In F. R. Volkmar, R. Paul, A. Klin, & D. Cohen (Eds.), *Handbook of autism and pervasive* *developmental disorders: Diagnosis, development, neurobiology, and behavior* (pp. 335–364). John Wiley & Sons Inc.

Thurm, A., Farmer, C., Salzman, E., Lord, C., & Bishop, S. (2019). State of the field: Differentiating intellectual disability from autism spectrum disorder. *Frontiers in Psychiatry, 10*. https://www.frontiersin.org/article/10.3389/fpsyt.2019.00526

Tomchek, S. D., & Dunn, W. (2007). Sensory processing in children with and without autism: A comparative study using the short sensory profile. *American Journal of Occupational Therapy, 61*, 190–200. https://doi.org/10.5014/ajot.61.2.190

Torres, R., DeBar, R. M., Reeve, S. A., Meyer, L. S., & Covington, T. M. (2018). The effects of a video-enhanced schedule on exercise behavior. *Behavior Analysis in Practice, 11*, 85–96. http://doi.org/10.1007/s40617-018-0224-1

Tuchman, R., Cuccaro, M., & Alessandri, M. (2010). Autism and epilepsy: Historical perspective. *Brain and Development*, *32*, 709–718. https://doi.org/10.1016/j.braindev.2010.04.008

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*, 1–10. https://doi.org/10.1007/BF02291170

Uljarević, M., Frazier, T. W., Phillips, J. M., Jo, B., Littlefield, S., & Hardan, A. Y. (2020). Mapping the research domain criteria social processes constructs to the Social Responsiveness Scale. *Journal of the American Academy of Child & Adolescent Psychiatry*, *59*, 1252-1263.e3. https://doi.org/10.1016/j.jaac.2019.07.938

Vaughan, C. A. (2011). Review of Childhood Autism Rating Scale (2nd ed.). *Journal of Psychoeducational Assessment*, *29*, 489–493. https://doi.org/10.1177/0734282911400873

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika,* 41, 321-327. https://doi.org/10.1007/BF02293557

Velicer, W. F., & Jackson, D. N. (1990). Component analysis versus common factor analysis: Some further observations. *Multivariate Behavioral Research*, *25*(1), 97–114. https://doi.org/10.1207/s15327906mbr2501_12

Veness, C., Prior, M., Bavin, E., Eadie, P., Cini, E., & Reilly, S. (2012). Early indicators of autism spectrum disorders at 12 and 24 months of age: A prospective, longitudinal comparative study. *Autism*, *16*, 163–177. https://doi.org/10.1177/1362361311399936

Volker, M. A., Dua, E. H., Bergamo, N. M., Searle, E. K., Zeng, S., Lopata, C., Thomeer, M. L., Toomey, J. A., Nelson, A. T., Rodgers, J. D., McDonald, C. A. (2022). *Revisiting the factor structure of the Gilliam Autism Rating Scale – Second Edition in an autism spectrum disorder sample*. [Manuscript in preparation]. Department of Counseling, Educational Psychology, and Special Education, Michigan State University.

Volker, M. A., Dua,, E. H., Lopata, C., Thomeer, M. L., Toomey, J. A., Smerbeck, A. M., Rodgers, J. D., Popkin, J. R., Nelson, A. T., & Lee, G. K. (2016). Factor structure, internal consistency, and screening sensitivity of the GARS-2 in a developmental disabilities sample. *Autism Research and Treatment, 2016*, 1-13. https://doi.org/10.1155/2016/8243079

Volker, M. A., Thomeer, M. L., & Lopata, C. (2012). Pervasive developmental disorders. In A.S. Davis (Ed.)*, Handbook of pediatric neuropsychology* (pp. 501-535). Springer Publishing Company.

Vukićević, S., Đorđević, M., Glumbić, N., Bogdanović, Z., & Đurić Jovičić, M. (2019). A demonstration project for the utility of Kinect-based educational games to benefit motor skills of children with ASD. *Perceptual and Motor Skills, 126*, 1117–1144. http://doi.org./10.1177/0031512519867521

White, S. W., Keonig, K., & Scahill, L. (2007). Social skills development in children with autism spectrum disorders: A review of the intervention research. *Journal of Autism and Developmental Disorders, 37*, 1858–1868. https://doi.org/10.1007/s10803-006-0320-x

Wiggins, L. D., Bakeman, R., Adamson, L. B., & Robins, D. L. (2007). The utility of the social communication questionnaire in screening for autism in children referred for early intervention. *Focus on Autism and Other Developmental Disabilities, 22*(1), 33–38. https://doi.org/10.1177/10883576070220010401

Wiig, E. H., Semel, E., & Secord, W. A. (2013). *Clinical evaluation of language fundamentals* (5th ed.). Pearson.

Witwer, A. N., & Lecavalier, L. (2007). Autism screening tools: An evaluation of the social communication questionnaire and the developmental behaviour checklist–autism screening algorithm. *Journal of Intellectual and Developmental Disability, 32*, 179-187. https://doi.org/10.1080/13668250701604776

Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology, 9*, 79-94. https://doi.org/10.20982/tqmp.09.2.p079

Zane, E., Arunachalam, S., & Luyster, R. (2021). Caregiver-reported pronominal errors made by children with and without autism spectrum disorder. In Danielle Dionne and Lee-Ann Vidal Covas (Eds.), *Proceedings of the 45th annual Boston University Conference on Language Development* (pp. 845-859). Cascadilla Press.

Zeldovich, L. (2018, May 9). The evolution of 'autism' as a diagnosis, explained. https://www.spectrumnews.org/news/evolution-autism-diagnosis-explained/

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (2011). *Preschool Language Scales 5.* Pearson.

Zumbo, B. D., Gadermann, A. M., Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods, 6*, 21-29. https://doi.org/10.22237/jmasm/1177992180

Zwaigenbaum, L., & Penner, M. (2018). Autism spectrum disorder: Advances in diagnosis and evaluation. *BMJ*, k1674. https://doi.org/10.1136/bmj.k1674