EXPLORING SEX DIFFERENCES IN THE STRUCTURE OF THE ADOS-2 IN AN EARLY INTERVENTION SAMPLE

A Dissertation

Submitted to

the Temple University Graduate Board

In Partial Fulfillment

of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

by

Brigid Garvin

May 2021

Dissertation Committee Members:

Dr. Frank Farley, Temple University

Dr. Catherine Fiorello, Temple University

Dr. Joseph DuCette, Temple University

Dr. James Connell, Drexel University

Copyright by

Brigid Garvin 2021

ABSTRACT

Autism Spectrum Disorder (ASD) is diagnosed using the same criteria for males and females (e.g., DSM-5, ICD-10). Our understanding of ASD, including its etiology, symptom presentation, and prevalence has evolved significantly over time motivating several changes to the diagnostic criteria and the tools with which symptoms are measured. One aspect of ASD prevalence and presentation that has remained consistent overtime and across the world is the consistently disproportionate male-to-female ratio that is reported. This ratio may reflect a true difference in prevalence or suggest that females with ASD are under- or mis-diagnosed, as evidenced by an emerging body of literature suggesting differences in symptom presentation and related abilities across sex. Studies have reached varied conclusions regarding sex differences as measured by standardized autism assessment tools, for example, finding no significant sex differences in Total Scores across measures or finding fewer parent reported repetitive and restricted behaviors (RRBs) in females (Duvekot et al., 2017). Limitations in the current body of research include small sample sizes, inconsistencies in participant ages, and the use of mixed sets of measures with a strong reliance on parent report. Most studies within this body of research use one of three iterations of the Autism Diagnostic Observation Schedule (ADOS). Indeed, the ADOS and its revisions are considered the gold-standard assessment tools for ASD symptoms and in many studies and clinical practice are used almost exclusively to determine whether a child meets diagnostic criteria. Previous studies have explored the factor structure of the ADOS in its current iteration (i.e., ADOS-2) by domain area and between ASD and typically developing or clinical samples of participants (Bishop et al., 2016; Gotham et al., 2007; Gotham et al., 2008). However, to date, no published studies have explored the factor structure of the ADOS-2 across sex with a preschool age sample.

The current study examined the structure of the ADOS-2 through exploratory factor analysis to analyze algorithm items and determine if the Social Affect (SA) and Restricted and Repetitive Behavior (RRB) subscales demonstrate a different structure for males and females. A two-factor structure, relatively consistent with the tool itself, was identified for the combined and male samples. However, the female sample produced a three-factor structure, with some loadings inconsistent with the current algorithm of the ADOS-2. Further analysis through confirmatory factor analysis indicated a relatively well-fitting model for the combined, male, and female groups. However, despite adequate model fit, there were notable differences in the strength at which certain items loaded onto each subscale across sex. Similarly, correlation estimates between the two subscales were stronger for the combined and male groups, and relatively weak for the female group. Implications for the use of the tool with females are discussed, along with suggestions for future evaluation of the ADOS-2 algorithm across sex.

TABLE OF CONTENTS

Page

ABSTRACT……...iii

LIST OF TABLES…….….vii

LIST OF FIGURES……...viii

CHAPTER

1. INTRODUCTION……..…1

Autism as a Psychiatric Condition in the DSM…….…1

Autism Spectrum Disorder: Etiology, Prevalence, and Presentation…….…...4

Questioning the Sex Ratio…….……7

2. REVIEW OF LITERATURE…….…...9

A Review of Bias in Psychological Assessment…….…..9

A Review of Potential Bias in ASD Criteria and Assessment…….13

Clinical versus Population Based Sampling……13

The Female versus Male Phenotype……14

Evaluating Diagnostic Measures……..15

The Current Study……21

3. METHODS……...……...23

Participants……...23

Measures……..24

Clinicians and Training……28

Setting……..28

Page

Procedures……29

4. RESULTS…….....32

Descriptive Statistics of Sample……...32

Correlation Analysis ……....33

Exploratory Factor Analysis……38

Confirmatory Factory Analysis……44

5. DISCUSSION……..52

Findings from Descriptive Analyses……54

Findings from Factor Analyses……57

Exploratory Factor Analysis: Combined and Male Groups……59

Exploratory Factor Analysis: Female Group……...62

Confirmatory Factor Analysis……63

Implications for Practice and Future Directions of Study …… …….67

Limitations ……..…….69

REFERENCES…….……71

LIST OF TABLES

Table Page

3.1 ADOS-2, Module One Algorithm……...28

4.1 Age at Evaluation and Diagnosis by Sex……32

4.2 Mean, Standard Deviation, Skewness, and Kurtosis of Each Group……..33

4.3 Bivariate Correlations Between Items for the Combined Group……35

4.4 Bivariate Correlations Between Items for the Male Group……36

4.5 Bivariate Correlations Between Items for the Female Group……37

4.6 Tests of Sample Adequacy for Combined, Male, and Female Groups……..39

4.7 Factor Loadings for Combined Group…….....41

4.8 Factor Loadings for Male Group……42

4.9 Factor Loadings for Female Group……43

4.10 Statistics for Model Fit for Combined, Male, and Female Groups……..44

4.11 Parameter Estimates for Latent Variables in the Model: Combined Group……46

4.12 Parameter Estimates for Latent Variables in the Model: Male Group…….46

4.13 Parameter Estimates for Latent Variables in the Model: Female Group…….....47

4.14 Order of Item Loadings for Male and Female Groups……51

LIST OF FIGURES

Figure Page

3.1 Structural Model Used for CFA……..…31

4.1 Structural Model with Standardized Coefficients: Combined Group……48

4.2 Structural Model with Standardized Coefficients: Male Group……49

4.3 Structural Model with Standardized Coefficients: Female Group……....50

CHAPTER

INTRODUCTION

Autism spectrum disorder (ASD) is a neurodevelopmental disability characterized by impairments in social communication and social reciprocity, in the presence of patterns of repetitive and restricted behavior (APA, 2013). Since its initial conceptualization as a form of childhood schizophrenia in the early 1900's, autism as a construct has evolved significantly over time, leading to several iterations of diagnostic criteria. Concurrent to the evolution of the criteria, ongoing lines of research have examined the etiology of the disorder, its prevalence, and the effectiveness of different treatment approaches. While the study of each of these aspects of ASD is usually designated to specific fields, for example, genetics, epidemiology, and the clinical or applied fields of psychology, behavior analysis, and education, the findings have a profound effect on one another leading to recommendations for clinical practice and policy that are inextricably linked. In order to fully understand the construct of autism spectrum disorder as we know it today, it is critical to consider the interaction between the evolution of the diagnostic criteria, changes in the prevalence of the disorder, and the progression of standardized diagnostic tools with which symptoms are measured.

Autism as a Psychiatric Condition in the DSM

The story of autism includes an evolving nosology that begins long before the publication of the first edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM) in 1952 and even before the well-known *Father of Autism*, Leo Kanner, identified the core features of the disorder. The description of *autistic traits* within the fields of psychiatry and psychology was first coined in 1910 by Swiss psychiatrist Eugen Bleuler who is credited as the *Father of Schizophrenia* (Verhoeff, 2013). Bleuler used the term to describe what was considered as one of the most notable symptoms of schizophrenia: a withdrawal from the external world (Verhoeff,

2013). According to Bleuler, *autistic thinking* was characterized by infantile wishes to avoid unsatisfying realities and replace them with fantasies and hallucinations (Evans, 2013). Later, Leo Kanner, an Austrian-American psychiatrist and physician, described a childhood version of schizophrenia to include extreme *autistic aloneness*, *delayed echolalia*, and *an anxiously obsessive desire for the maintenance of sameness* (Evans, 2013). In keeping with this perspective, the first and second editions of the DSM (1952 and 1968) defined autism as a form of infantile schizophrenia marked by a detachment from reality that was observed at a particularly early age (Volkmar, 1998). It wasn't until the publication of the DSM-III in 1980 and its revision in 1987, influenced by the early work of researchers like Rutter (1978) and Wing (1981), that autism was established as its own separate diagnosis and distinct from schizophrenia (Volkmar, 1998).

In addition to finally differentiating autism and schizophrenia, the DSM-III presented another nosologically important change to the way symptoms were described. This version of the diagnostic criteria highlighted a set of symptoms that were more objective than previous iterations, with specific behaviors or characteristics included to help guide practitioners through their decision-making and diagnostic process (APA, 1980). It was also the third iteration of the DSM that conceptualized symptoms in three domains: a lack of interest in people, severe impairments in communication, and *bizarre* responses to the environment, importantly all of which were observed to develop in the first 30 months of life (Volkmar et al., 1992). The revision of the DSM-III in 1987 once again significantly altered the autism criteria. It broadened the concept of autism by adding a diagnosis at the mild end of the spectrum: pervasive developmental disorder-not otherwise specified (PDD-NOS) (APA, 1987). Even though the manual did not yet include the word *spectrum*, this change reflected the growing understanding

among researchers and clinicians that autism was not a single condition but rather a group of related conditions that emerge early in development and remain present throughout life (Volkmar et al., 1992). Adding PDD-NOS allowed clinicians to include children who did not fully meet the criteria for autism but still required developmental or behavioral support. The DSM-IV, released in 1994 and revised in 2000 (DSM-IV-TR), was the first edition to categorize autism as a *spectrum* of disorders. However, this version continued to present a set of five separate conditions with distinct features (APA, 1994). In addition to autistic disorder and PDD-NOS, it added Asperger's disorder, also at the mild end of the spectrum; childhood disintegrative disorder (CDD), characterized by severe developmental regressions; and Rett syndrome, which developed primarily in females and included a degenerative motor component (APA, 1994). This breakdown and the inclusion of Rett syndrome, which had a presumed genetic etiology, reflected the popular hypothesis at the time that autism itself was rooted in a genetic difference, and that each of these individual disorders would ultimately be linked to a set of identifiable genetic markers (Witwer & Lecavalier, 2008).

There were several clinical and empirical findings following the issue of the DSM-IV and DSM-IV-TR that served as the impetus for yet another change to the diagnostic criteria. Over time, it became clear that finding isolated genetic makers for autism and its related disorders was unlikely (Sanders et al., 2015). Other misconceptions about the distinct nature of each disorder included the expectation of different long-term outcomes and the appropriateness of separate treatments for autistic disorder, Asperger's disorder, and PDD-NOS (Szatmari, 2000). The lack of evidence to this claim and the inability of clinicians in the field to reliably differentiate these disorders diagnostically (Snow & Lecavalier, 2011; Walker et al., 2004; Witwer & Lecavalier,

2008) led to yet another revision of the DSM, which would characterize autism as an all-inclusive diagnosis.

The DSM-5 introduced the APA's current criteria for autism spectrum disorder in 2013. This diagnosis is characterized by two groups of features present in early childhood: (a) persistent impairment in reciprocal social communication and social interaction and (b) the presence of restricted and repetitive patterns of behavior (APA, 2013). The current iteration eliminated Asperger disorder and PDD-NOS and added the diagnosis of Social Communication Disorder to include children with only language and social impairments. Childhood disintegrative disorder and Rett syndrome were removed from the autism category altogether (APA, 2013).

Autism Spectrum Disorder: Etiology, Prevalence, and Presentation

While the nuances of symptom presentation in the context of an *autism spectrum* have generally gained acceptance among researchers and clinicians, other aspects of the disorder, namely its etiology, have remained less clear despite decades of rigorous scientific investigation. Rather than focusing on an isolated genetic cause as previously hypothesized, the general consensus in the scientific community indicates the greater likelihood of an interaction between one's genes, biology, and the environment that make a child more likely to have an ASD (Centers for Disease Control and Prevention, 2020). Indications of a strong genetic influence in the development of some cases is elucidated through twin and sibling studies which demonstrate a 70% - 90% concordance rate among monozygotic twins, approximately 30% among dizygotic twins, and approximately10% among non-twin siblings (Bai et al., 2019; Constantino et al., 2013; Rosenberg et al., 2009). In addition, individuals with specific genetic or chromosomal conditions, such as fragile X syndrome, Down syndrome, or tuberous sclerosis, demonstrate

higher rates of ASD as compared to the general population (Cohen et al., 2005; Sztainberg et al., 2016). Further still, research based in examinations of the human exome has identified hundreds of corresponding genetic markers that have a weak link to risk for ASD and approximately 65 that have a slightly stronger link (Sanders et al., 2015). Other research has identified the risk of biological and environmental factors, such as prenatal exposure through maternal use of specific prescription drugs (e.g., valproic acid and thalidomide) (Christensen et al., 2013; Stromland et al., 1994), low birth weight (Pinto-Martin et al., 2011), advanced parental age, and birth complications that cause ischemia and hypoxia (Modabbernia et al., 2017). Still, other related factors such as maternal obesity, maternal diabetes, and the occurrence of a caesarian section have shown weaker yet significant associations with ASD risk (Modabbernia et al., 2017). Further, in refining the potential causal variants, recent studies have demystified some commonly discussed factors such as exposure to vaccination, maternal smoking, and the use of assisted reproductive technologies which are found to be unrelated to risk of ASD (Modabbernia et al., 2017).

Research will likely continue to reveal new findings about these and perhaps new etiologies. Other ongoing investigations examine the prevalence of the disorder within the United States and world-wide. These data suggest prevalence rates that have changed drastically over time. In 2002 the Centers for Disease Control (CDC) established the Autism and Developmental Disabilities Monitoring (ADDM) Network to systematically track the number and characteristics of children with ASD in multiple communities in the United States. Data gathered through the ADDM Network demonstrate a consistent increase in prevalence across communities, ranging from 1 in 150 children in 2000 to the most recent estimate of 1 in 54 children in 2016 (Maenner et al., 2020). Even within the most recent US sample, variability in

prevalence exists across states, with 1 in 76 children identified in Colorado and 1 in 32 identified in New Jersey (Maenner et al., 2020). These data also note changes in related abilities or characteristics of ASD. For example, one third of children with ASD in the CDC sample experienced a comorbid intellectual disability; while a decade ago it was thought that approximately two thirds of children identified with ASD had significantly below average cognitive functioning (Centers for Disease Control and Prevention, 2020). Prevalence data also demonstrate that ASD occurs at near equal rates among all racial, ethnic, and socioeconomic groups (Centers for Disease Control and Prevention, 2020).

Notwithstanding the considerable variability across and within geographical regions, estimates of the prevalence of autism collected since the year 2000 in North America, the Western Pacific (Japan and China), and Europe do not differ statistically. Generally, reviews and metanalyses that consider studies conducted since the 1960s suggest that there has been a consistent rise in prevalence over time across the world (Elsabbagh et al., 2012). Such time trends have given rise to the hypothesis that there is indeed a secular increase in rates of autism. Still, some researchers and clinicians have suggested that this rise is due in large part to the refinement of the diagnostic criteria. In response, studies have evaluated the effect of changes in diagnostic criteria and how they lend to changes in prevalence. While these studies conclude with mixed results, most comprehensive reviews suggest that only one third of the rise in prevalence can be attributed to changes in the criteria (King & Bearman, 2009).

Prevalence rates have evolved over time, likely as a result of a still unknown environmental, genetic, and biological interaction in combination with a refined clinical interpretation of symptoms. Our ability to measure these symptoms has, in turn, been influenced by revisions to the diagnostic criteria and the subsequent development of diagnostic tools to align

with that criteria. However, no matter how well developed our tools and clinical acumen have become, differences in prevalence across sex is one feature of the disorder that has been consistently revealed over time. The most recent CDC data indicate that males are four times more likely to be identified with ASD than females in the US (Maenner et al., 2020). Furthermore, despite some variability among epidemiologic studies world-wide, international analyses consistently indicate higher ratios of ASD in males relative to females (Elsabbagh et al., 2012). This strikingly consistent feature has motivated influential ideas about the nature and etiology of ASD, such as the extreme male brain (Baron-Cohen, 2002), the female protective genetic effect (Jacquemont et al., 2014), the ability of females to "camouflage" symptoms (Dean et al., 2017), and female autism phenotype theories (Hull et al., 2020; Kopp & Gillberg, 1992).

Questioning the Sex Ratio

While female protective biological factors such as those proposed by Jacquemont and colleagues (2014) are shared by many in the scientific community, there is also evidence suggesting a bias in the assessment and diagnostic processes which strongly influences the likelihood that males receive a diagnosis more often and earlier than females (Loomes et al., 2017). In particular, studies have found that females often require the presence of more significant and a greater number of behavioral difficulties than males to receive a diagnosis, despite having equivalent levels of autism symptoms (Duvekot et al., 2017; Dworzynski et al., 2012; Shattuck et al., 2009). In addition, females who receive an autism diagnosis do so at a later age than males on average (Begeer et al. 2013; Kirkovski et al. 2013; Rutherford et al. 2016). Further, studies evaluating the social behavior of school-age females and males with autism as compared to their non-ASD peers indicate that females with a diagnosis of ASD present with less obvious, or camouflaged, social differences when observed in the context of the natural

social landscape of typically developing children (Dean et al., 2017). One way to begin to dissect this complicated disparity and explore the possibility of assessment bias, is to analyze our psychodiagnostics tools and the diagnostic process.

CHAPTER 2

LITERATURE REVIEW

A Review of Bias in Psychological Assessment

The issue of bias in diagnostic criteria and psychological assessment remains controversial and has been widely debated (Jane et al., 2007). Conversely, physical and biological measurements such as height, weight, and even risk of heart disease elicit little controversy, although results of these assessments vary across sex, cultural, racial, and ethnic groups (Reynolds & Suzuki, 2012). As explained by Reynolds and Suzuki (2012), this is true in part because measurements of physical and biological traits are absolute, obtained and verified in direct and relatively simple ways, and free from the distinctive social implications and consequences of standardized test scores. It is the nature of psychological characteristics and their measurement that is partly responsible for long-standing concern over test bias (Reynolds & Suzuki, 2012). Simply stated, psychological characteristics are internal, meaning scientists cannot observe or measure them directly and, instead, infer them from external behavior.

Perhaps the most robust literature base in psychological test bias explores the question of biases within measures of cognition. Suggestions of cultural, racial, and ethnic biases in assessments of cognitive ability have attracted public attention and controversy since the early 1900s with the publication of Binet's first intelligence scale (Reynolds & Suzuki, 2012). The implications of potential bias in measures of intelligence have even motivated some states to develop legislation placing restrictions on their use with specific populations. When considering bias related exclusively to sex, most reviews highlight the differences in prevalence of psychopathologies within groups of personality and mood disorders (Jane et al., 2007). There are many ways to interpret differential prevalence rates as a function of sex, including the

consideration that the criteria for certain mental and behavioral health disorders assume unfairly that stereotypical female or male characteristics are pathological, depending on the disorder (Jane et al., 2007). For example, Lindsay and Widiger (1995) found that items in several of the most widely used instruments to assess personality disorders are biased in the sense that they are endorsed more easily by males than females. Widiger (1998) went on to describe specific ways in which sex bias may be related to differential prevalence rates for psychological disorders, including: (a) biases in diagnostic constructs, (b) inaccurate diagnostic thresholds, (c) biases in sampling procedures, (d) misapplication of the diagnostic criteria, and (e) biased assessment instruments. Of these possible explanations, greater support has emerged for the argument that assessment instruments may contain sex bias, and clinicians may behave in a biased way when they apply certain criteria to females and males (Jane et al., 2007).

In statistics, *bias* refers to systematic error in the estimation of a value. A biased test is one that systematically overestimates or underestimates the value of the variable it is intended to assess. If this bias occurs as a function of a nominal cultural variable, such as ethnicity or gender, cultural test bias is said to be present (Reynolds & Suzuki, 2012). Before assessing the presence or kind of bias in an established measure, it may be helpful to review the process of test development. Test development procedures are essentially the same for all standardized tests (Ramsay & Reynolds, 2000). Ramsay and Reynolds (2000) describe a process which begins with developing or collecting a large pool of items thought to measure the characteristic of interest. Theory and practical usefulness are standards commonly used to select an item pool. Ramsay and Reynolds describe the selection process as a rational one, as it largely depends on reason and judgment. A rigorous means to carry out the item selection process at this stage simply does not exist. Therefore, test authors have no generally accepted evidence that they have selected

appropriate items. A natural second step includes discarding items of suspect quality, again on rational grounds, to reduce the pool to a manageable size. Next, the author administers the items to a group of examinees. Statistical procedures then help to identify items that seem to be measuring an unintended characteristic, usually causing the author to discard or modify those items. Finally, examiners administer the remaining items to a large, diverse group of people called a standardization sample or *norming sample*. This sample should reflect every important characteristic of the population that will take the final version of the test, which for a test intended for the general population, should include relatively equal numbers of males and females. Statisticians compile the scores of the norming sample into an array called a *norming distribution*. In order to address concerns regarding group representation in the norming sample, some test developers engage in group oversampling (i.e., including larger numbers of individuals from a particular group above and beyond their proportional representation in the overall population). Supplemental norms may be created for a particular group thereafter. It is usually the lack of attention to specific group representation and the underuse of oversampling that motivate charges of bias. For example, in the case of racial and ethnic biases, it is suggested that low proportions of minorities in the norming sample of a test correspondingly lead to influences on test results (Ramsay & Reynolds, 2000).

While efforts to prevent test bias and explore reasons for group differences are critical, other researchers caution quickly made assumptions about bias in the absence of rigorous scientific evaluation. Jensen (1980) and Brown and colleagues (1999) contended that three fallacious assumptions were impeding the scientific study of test bias: (1) the *egalitarian fallacy*, that all groups were equal in the characteristics measured by a test, so that any score difference must result from bias; (2) the *culture-bound fallacy*, that reviewers can assess the culture

loadings of items through casual inspection or armchair judgment; and (3) the *standardization fallacy*, that a test is necessarily biased when used with any group not included in large numbers in the norming sample. Taking Jensen's advice into account, questions about bias in tests should be explored rigorously and with recognition of these fallacious assumptions.

Jensen (1980) suggested that test bias impacts two forms of validity: predictive and construct validity. Bias in predictive validity, as defined by Jensen, is systematic error in predicting a criterion variable for people of different groups. This bias occurs when one regression equation is incorrectly used for two or more groups. In previous research examining test bias in predictive validity, partial correlation procedures (Reynolds et al., 1984; Wilson et al., 1989), chi-square techniques (Jensen, 1976), and path analysis (Keith & Reynolds, 1990; Ramsay, 1997) have been used. Bias in construct validity occurs when a test measures groups of examinees differently. For example, a test can be more difficult, valid, or reliable for one group than for another. Construct bias involves the test itself, whereas predictive bias involves a test's prediction of a result outside the test (Jensen, 1980). Anastasi (1988) further defined construct validity as the extent to which a test may be said to measure a theoretical construct or trait. Test bias in construct validity, then, may be defined as the extent to which a test measures different constructs for different groups. Factor analysis is a widely used method for investigating construct bias (Reynolds, 2000). This set of complex techniques groups together items or subtests that correlate highly among themselves. When a group of items correlates highly together, the researcher interprets them as reflecting a single characteristic. The researcher then examines the pattern of correlations and induces the nature of this characteristic. Very similar factor analytic results for two or more groups, such as sex or race, are evidence that the test responses being analyzed behave similarly as to the constructs they represent and the extent to

which they represent them. As noted by Reynolds and colleagues (1999), such comparative factor analyses with multiple groups are important for the work of clinicians, who must know that a test functions very similarly from one group to another to interpret scores consistently.

A Review of Potential Bias in ASD Criteria and Assessment

Many theorists and researchers highlight the fact that behavioral markers used as diagnostic criteria have been long-established based on autism symptom presentation in samples which are overwhelmingly male (Volkmar et al., 1994). Subsequently, the tools used to measure symptoms have been developed based on predominantly male populations (Kirkovski et al., 2013; Kopp & Gillberg, 2011; Mattila et al., 2011). Therefore, it is possible that the diagnostic criteria lack sensitivity to these potential sex differences in the manifestation and presentation of ASD. That is, when females are evaluated against these criteria and with these tools, they may be less likely to receive a diagnosis of ASD even when clinically significant characteristics of the disorder are identified (Wilson et al., 2016).

Clinical versus Population Based Sampling

Most research that questions sex ratios also critiques conclusions that are made from clinical samples of individuals with an already established diagnosis of ASD. These samples consistently include more males than females and thus underestimate the number of females that may be present in the population at large (Loomes et al., 2017). These critiques highlight another fundamental bias within sampling procedures and call attention to where and how sex ratios are established. Russell and colleagues (2011) examined a sample of cases using both diagnostic and trait measures of ASD to determine what distinguishes females who meet diagnostic criteria from those who do not and if the same distinguishing features were apparent in males. Their results indicated that a diagnosis was found to be more likely for males than for females even

when severity of symptoms was held constant. A review by Loomes and colleagues (2017) demonstrated that more females have been identified in studies that actively sought cases in the general population, regardless of whether the child had already been identified by clinical or educational services as compared to studies that passively evaluated referred or high-risk samples. These results have led researchers to hypothesize that there are likely females in the general population who, if assessed, would meet criteria for ASD, but who are not referred in practice and therefore never receive a clinical diagnosis (Dworzynski et al., 2012; Loomes et al., 2017). In summary, this research suggests that females with autism are at greater risk than males of having their ASD overlooked, misdiagnosed, or diagnosed late (Loomes et al., 2017).

The Female versus Male Phenotype

As questions arise about the true sex ratio across studies, researchers have investigated whether males and females present with clinically different symptom phenotypes. Evidence has been mixed with respect to these differences as they relate to the core symptoms of ASD. Some studies indicate that males have greater social and communicative challenges compared with females (Beggiato et al., 2017; Hiller et al., 2016), whereas others show the opposite pattern (Carter et al., 2007; Frazier et al., 2014; Hartley & Sikora, 2009), and still others show no particular differences in social communication skills (Mandy et al., 2012; Reinhardt et al., 2015; Szatmari et al., 2012). Similarly, some studies have found males with ASD to exhibit higher levels of repetitive and stereotyped behaviors than females (Beggiato et al., 2017; Hartley & Sikora, 2009; Szatmari et al., 2012), whereas others have found no differences in this domain (Carter et al., 2007; Holtmann et al., 2007; Reinhardt et al., 2015). In their metanalysis, Van Wijngaarden-Cremers and colleagues (2014) found that among individuals with ASD beyond age six, males show higher levels of restricted and repetitive behaviors and interests than

females; no significant sex differences were found for social interaction or communication (Van Wijngaarden-Cremers et al., 2014).

Evaluating Diagnostic Measures

One way to begin to dissect this complicated disparity in sex ratio is to analyze the validity and reliability of the ways we measure autism symptoms; that is, evaluating tools that are considered the gold standard in autism assessment across sex. Tools like the Autism Diagnostic Observation Schedule (ADOS) (Lord et al., 1999) and its subsequent revision (Lord et al., 2012) have undoubtedly influenced the standard of assessment for ASDs in a positive manner by providing a tool standardized in both administration and scoring procedures. The structure offered through this tool makes it an essential choice for use in research and clinical practice alike (McCrimmon & Rostad, 2014).

The ADOS has evolved from a series of tools used primarily in research, including the Autism Diagnostic Observation Schedule (Lord et al., 1989), the Pre-Linguistic Autism Diagnostic Observation Scale (PL-ADOS: DiLavore et al., 1995), and the Autism Diagnostic Observation Schedule, Generic (Lord et al., 2000), which is the designated name of the tool prior to its formal publication. The evolution of the original version in 1989 to the PL-ADOS and ADOS-G were motivated by efforts to create a tool appropriate for use with a wide range of young children and adolescents, including those with minimal language and limited ability to attend to structured tasks such as conversation and tabletop activities (Lord et al., 2012). The Autism Diagnostic Observation Schedule (Lord et al., 1999), was eventually published for use in the clinical community through Western Psychological Services. Finally, a revised version was published in 2012, the Autism Diagnostic Observation Schedule, Second Edition (Lord et al., 2012). There are five modules included in the ADOS-2 and each module consists of specified

tasks designed to elicit communicative, social, and behavioral characteristics consistent with ASD. The decision about which module to administer to the examinee is dependent on the individual's age and language level. Behavioral and language observations are noted throughout the administration, particularly regarding social and communicative behaviors related to ASD diagnostic criteria. There are several items to code following the administration, ranging from 29 to 41 across modules, though only specific items contribute to the final algorithm, which produces a score describing the presence of symptoms in the examinee as compared to individuals who meet criteria for ASD (Lord et al., 2012).

While the diagnostic criteria presented in the DSM-III R and DSM-IV-TR at the time of the development of the original versions of the tool (e.g., ADOS, PL-ADOS, and ADOS-G) presented symptoms across three domains; including social impairments, language impairments, and repetitive and restricted behaviors, the algorithms did not originally include items related to repetitive behaviors and restricted interests. Only two domains comprised the algorithm, including items measuring social and communication abilities. Initially, this choice was based on suggestions that the context of the assessment was too limited in time to observe repetitive and restricted behaviors comprehensively (McCrimmon & Rostad, 2014).

Over time, studies sought to refine the sensitivity and specificity of the ADOS algorithms in response to concerns with the association between the social domain total scores and level of cognitive impairment in preschool children (Joseph et al., 2002), the overidentification of children with specific language impairments (Bishop & Norbury, 2002), the tool's limited validity in adequately identifying ASD in young children with mild intellectual disabilities (de Bildt et al., 2004), and the inadequate identification of nonverbal individuals with autism spectrum disorders on the Model One algorithm specifically (Gotham et al., 2007). In 2007,

Gotham and colleagues proposed revised algorithms to improve the sensitivity and specificity of the ADOS. In the revised algorithms, the original domains were collapsed into a single factor consisting of items that describe social *and* communication symptoms, calling it the Social Affect factor (SA). In addition, 4 items from a second factor, Restricted and Repetitive Behavior (RRB), were included because ongoing research demonstrated that it was in fact possible for RRB domain items to contribute to a diagnosis, even in the limited context of the ADOS (Lord et al., 2006). Gotham and colleagues (2007) confirmed the two-factor model through exploratory and confirmatory factor analysis. While a third factor was produced which included items measuring joint attention, gesturing, and eye contact, the authors concluded that the two-factor model was more consistent across Modules One and Two and more parsimonious (Gotham et al., 2007).

Subsequent studies replicated the revised ADOS algorithms proposed by Gotham and colleagues (2007). In the first, small study ($N = 26$), Overton and colleagues (2008) reported slightly more accurate results with the revised algorithm for Module One, though results for Modules Two and Three were similar to those for the original algorithms. Gotham and colleagues (2008) replicated their own study in a large and independent sample ($N = 1282$) and presented sensitivity and specificity data that were similar or better than those for the original algorithms, with the exception of scores for young children with phrase speech and a clinical diagnosis of pervasive developmental disorder not otherwise specified (PDD-NOS). In another replication study ($N = 195$), Gray et al. (2008) found improved sensitivity for Module One with the revised algorithm. In yet another study, De Bildt and colleagues (2009) evaluated the revised algorithms in a Dutch sample ($N = 558$), finding the sensitivity and specificity of Module One to be modestly improved. While the overall consensus of Gotham and colleagues (2007 and 2008)

and the subsequent replication studies pointed to improvements in the tool's psychometric properties with the revised algorithm, none of these studies reported evaluations of the revised algorithm's fit across sex and all of these studies included predominantly male samples.

When the second edition was published (ADOS-2), it presented a revised algorithm including a Social Affect domain and a Restricted and Repetitive Behavior domain, similar to the model suggested by Gotham and colleagues (2007). Upon revision, the authors provided data on the validation sample of 1,630 distinct assessments primarily from sites in Illinois and Michigan. The sample included participants who were 57%-86% male and 71-91% white. The final data set included 56% of cases with a diagnosis of Autistic Disorder, 27% meeting criteria for PDD-NOS or Asperger's Disorder (i.e., non-autism spectrum), and 17% non-spectrum cases. Within the non-spectrum group, 41% had nonspecific intellectual disabilities, 21% had language disorders, 14% had Down Syndrome, 12% had Attention Deficit Hyperactivity Disorder or Oppositional Defiant Disorder, 6% had a mood or anxiety disorder, and 6% had an unspecified early delay (Lord et al., 2012). The replication sample included greater geographic diversity, with 1,282 distinct assessments for children across 10 sites throughout the United States (e.g., Boston, Washington, Utah, New York, California, Colorado, and Maryland). Sample demographics included participants who were 72% male and 91% white. However, clinical results of the sample noted less diversity with 75% of cases meeting criteria for Autistic Disorder, 7% meeting criteria for non-autism spectrum (PDD-NOS or Asperger's Disorder), and 17% with non-spectrum conditions (Lord et al., 2012). An exploratory multifactor item-response analysis was used to organize the codes into new domains: SA and RRB. The two-factor solution and goodness of fit were confirmed following further exploratory and confirmatory factory analyses. The authors specifically noted that the item *Stereotyped and Idiosyncratic Use of Words or*

Phrases represented an item in the communication domain on the original ADOS (Lord et al., 1999) algorithm, though results of the exploratory factor analysis demonstrated its strong loading to the RRB domain of the new algorithm, thus placing it there in the final revision (Lord et al., 2012). Goodness-of-fit results following the confirmatory factor analysis indicated a Comparative Fit Index of 0.97, Root Mean Square Error Approximation of 0.09 for Module One specifically (Lord et al., 2012). For Modules One through Three, logistic regressions indicated that both the SA and RRB domains made significant independent contributions to the prediction of diagnosis. However, the overall total score produced the highest predictive value, supporting the use of this score in diagnostic decisions. Results of internal consistency indicated Cronbach's α values for Modules One through Three as high for the SA domain (0.87-0.92) and moderate for the RRB domain (0.51-0.66). Test–retest reliability was calculated for Modules One through Three using a sample of 75 participants with analyses indicating that SA, RRB, and overall total scores had correlations ranging from 0.68 to 0.92. Receiver operating characteristic curves evaluated the predictive validity of the ADOS-2 to identify individuals on the autism spectrum accurately. When comparing the ADOS to the ADOS-2, sensitivity and specificity values were largely comparable or improved with the new algorithms, particularly in Modules One and Two (McCrimmon & Rostad, 2014).

While the disproportionate male to female ratio in studies that established and replicated the psychometric properties of the tool were critiqued, these samples certainly reflected the well-accepted sex disparity in the population at large (Lord et al., 2000; Gotham et al., 2007; Gotham et al., 2008; Oosterling, et al., 2010). Later studies evaluated the sensitivity and specificity of the ADOS-G, ADOS, and ADOS-2, as well as their concurrent validity with other measures of autism symptoms, in relation to male and female symptom presentation. Mussey and colleagues

(2017), in their investigation of a large clinical sample, suggested that the gold-standard measures for autism may be underestimating symptom severity in females. They discovered that whereas males and females who met criteria for ASD score similarly on the Childhood Autism Rating Scale (CARS), females scored lower on the Autism Diagnostic Observation Schedule–Generic (ADOS-G), indicating less autism symptoms. Similarly, when comparing males and females within a mixed sample of young children at high and low risk for ASD, Zwaigenbaum and colleagues (2012) found that females evidenced slightly fewer symptoms of ASD using the Autism Diagnostic Observation Scale Calibrated Severity Scores (Gotham et al., 2009) as well as fewer social and communication symptoms on the Autism Diagnostic Interview-Revised (Lord et al.,1994), which is the widely used interview counterpart to the ADOS. It is important to remind the reader that that the ADOS-G and ADOS did not yet include an RRB symptom domain.

While Ratto and colleagues (2018) did not conclude significant differences across sex when comparing total scores derived from the ADOS-2 and its interview counterpart (ADI-R) they suggested an interesting hypothesis considering the differences that *were* found in their sample on tools that rely on sex-based norms. The authors suggested that tools such as the Vineland-II and the Social Responsiveness Scale, which use sex norms to calculate subdomain and total scores, were more sensitive to possible sex differences, whereas tools like the ADOS-2 and ADI-R, which are best conceptualized as symptom-count measures developed on predominantly male samples without sex-specific cutoffs or algorithms, may not be. In conclusion, Ratto and colleagues questioned the sensitivity of gold standard autism diagnostic tools for females. They further argued that the variability in the sex ratio found in their study by cognitive ability, which was two males to one female among those with intellectual disability

(ID) and as high as six males to one female among those without ID, also raised the question of whether females with higher cognitive abilities were missed. Similarly, they found females as less likely to meet ADI-R criteria as intellectual functioning increased, which was not the case for the male sample (Ratto et al., 2018). There is further suggestion in the results of studies such as these that the over-reliance on gold standard measures in research trials creates a self-perpetuating cycle by which females whose autism symptoms may not align with the ADOS and ADI-R are excluded from research. Researchers then fail to find sex differences in their studies because the very females who may be most interesting to study are not included (Ratto et al., 2018).

The Current Study

There remains a significant gap in the literature comparing the structure of the ADOS-2 across sex, specifically in samples that include a large number of young participants, reflecting the age at which most children are diagnosed with ASD and therefore the age at which a potential diagnosis in females is overlooked. The current study, a secondary dataset analysis, included two primary aims. First, the structure of the ADOS-2, Module One algorithm was explored for the combined sample and across sex to examine the underlying processes that could produce correlations among each variable. This aim intended specifically to determine if the male and female samples yielded a two-factor model as indicated by the tool – Social Affect (SA) and Restricted and Repetitive Behavior (RRB) – or if a different structure was suggested across sex. Second, a confirmatory factor analysis was conducted to determine if the correlations among variables were consistent with the hypothesized structure for the combined sample, and then for males and females as two distinct groups. Subsequent analyses compared the fit of the

two-factor model across sex. An analysis of the descriptive statistics was also conducted for the combined, male, and female samples.

CHAPTER 3

METHODS

Participants

Participants included 736 children who were referred to a University clinic for a psychological evaluation by their early intervention team. All participants were previously deemed eligible for preschool-age early intervention services within a large, urban setting. As per federal and state guidelines, eligibility for early intervention included the presence of a delay greater than or equal to 25% in one or more of the following domains as measured through standardized developmental inventories: Cognitive Development, Communication Development, Motor Development, Social-Emotional Development, or Adaptive Behavior Development. The developmental inventories used to determine eligibility included the Battelle Developmental Inventory, Second Edition or the Developmental Assessment of Young Children, Second Edition. Specific demographic information for this sample was not available as part of the retrospective analysis, however, demographic data are provided on the school-age population in the urban setting for reference. The school-age population reported 52% of students as Black; 21% as Latinx; 14% White; 7% Asian; and 5% Multiracial/Other. Reports from the state's department of education indicates that 69% of student families in the district are considered low-income or economically disadvantaged.

Participants included children who met criteria for a diagnosis of ASD as well as those who did not meet criteria but who were deemed Developmentally Delayed (DD). Participants who did not meet criteria for ASD remained eligible for early intervention due to their developmental delays. Only children whose caregivers identified their primary language as either English or Spanish were included. Children with a primary language of English were evaluated

by an English-speaking clinician, and children with a primary language of Spanish were evaluated by a Spanish-speaking clinician.

Measures

Diagnoses of ASD were determined based on the current diagnostic criteria identified in the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5). Children who did not meet criteria for ASD were considered Developmentally Delayed, based upon the results of their previous early intervention evaluation. Evaluation measures included the Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) (Lord et al., 2012), a semi-structured developmental caregiver interview, and a parent rating scale. As described in previous sections, the ADOS-2 is a semi-structured, standardized measure of communication, social interaction, play, imagination, and restricted and repetitive behaviors published by Western Psychological Services. Used in clinical and research settings, the ADOS-2 is often referred to as the "gold standard" measure of observational assessment for autism spectrum disorder (McCrimmon & Rostad, 2014). It is a tool that can be administered and interpreted by appropriately credentialed professionals from medicine, psychology, or a related discipline in approximately 40 to 60 min, depending on the module selected and the specific behavior demonstrated by the examinee. Extensive prior experience with individuals with ASD is strongly recommended, and administrators must have completed formal clinical training. Each module consists of specified tasks designed to elicit communicative, social, and behavioral characteristics consistent with ASD. The format of the ADOS-2 differs from the majority of other standardized assessment measures in that the assessment is dynamic and examiners should not constantly refer to the manual during administration so that social interaction with the examinee is natural and maintained (McCrimmon & Rostad, 2014).

Administration necessitates delivery of specific tasks in a semi-structured manner whereby examiners deliberately modify their behavior according to a hierarchy of structured and unstructured social presses while observing the examinee's behaviors in response. Administration of tasks is flexible and does not have to follow a set sequence. Behavioral and language observations are noted throughout the administration, particularly regarding social and communicative behaviors related to ASD diagnostic criteria. The tasks are standardized in that examiners are provided instructions regarding task administration and the hierarchy of behaviors to be employed. Coding occurs directly following the administration using the behavioral notes taken by the examiner. Behavioral and language indicators noted during the assessment inform specific coded items for each module pertaining to (a) Language and communication, (b) reciprocal social interaction, (c) play and imagination, (d) stereotyped behaviors and restricted interests, and (e) other behaviors.

The ADOS-2 encompasses four modules, each with its own schedule of activities that allow clinicians to observe behavior in participants of particular developmental and language levels, ranging from those with no expressive language to verbally fluent children and adults. Only participants who received Module One of the ADOS-2 were included in the current study. Module One is administered to children over the age of 30 months whose verbal language abilities range from no verbal language to the use of single words or two-word utterances. Children with verbal language abilities characterized by phrased or fluent speech were administered either Module Two or Module Three of the ADOS-2, as per guidelines established in the manual, and not included in this study. Coding is a primary activity of the ADOS-2 and is the basis for clinical decision making when the codes are transferred to the algorithm and summed. As such, familiarity with coding concepts and conventions are integral to effective and

appropriate ADOS-2 administration. Specifically, a code of 0 indicates no evidence of abnormal behavior as specified by the code, 1 indicates mildly abnormal or slightly unusual behavior, 2 indicates definite abnormality of behavior, and 3 indicates markedly abnormal behavior. Other codes (7, 8, and 9) are used to identify specific language ability, atypical behaviors that are present but not of the specified type, or an item that cannot be scored for a number of possible reasons. On the Module One, the clinician rates the individual across a total of 34 items and scores from 16 possible items are selected and summed to generate a Social Affect score and a Restricted and Repetitive Behaviors score. Coding of items allows the examiner to distinguish severity in specific behaviors by rating a 3 as compared to a 2 and allowing for flexibility for a spoiled or difficult task or adaptations required to support diverse learners by using codes 7, 8, and 9. However, when using coded items to populate the algorithm, it is standard convention to convert scores of 3 to 2 and codes of 7, 8 , or 9 to 0. Algorithm items from the Social Affect and Restricted and Repetitive Behavior subscale scores are totaled to generate a Total Algorithm score, which is compared to a cutoff score to assist in determining diagnosis (Lord et al., 2012). See Table 3.1 for a list of the items included in the ADOS-2, Module One Algorithm.

Validation and Replication studies for the ADOS-2 reported internal consistency with Cronbach's α values for Module One ranging from 0.87-0.92 for the SA domain and 0.51-0.66 for the RRB domain. Test–retest reliability was calculated using a sample of 75 participants who were administered the same module twice within an average of 10 months. Analyses indicated that SA, RRB, and overall total scores had correlations ranging from 0.68 to 0.92. During validation, examiners had to achieve research reliability (i.e., 80% or greater exact agreement with other reliable raters). Across the five modules, interrater reliability for item coding was 71% or higher with the majority of mean weighted kappas exceeding 0.60. Correlations were

calculated for SA, RRB, and overall total and ranged from 0.79 to 0.98 across the five modules. Agreement in diagnostic classification ranged from 92% to 98% in Modules One through Three. Codes were selected for inclusion based on their unique contribution to the original ADOS algorithm (i.e., no interitem correlations greater than .70). A two-factor solution was identified based on exploratory and confirmatory factor analysis, comprising the SA and RRB domains. For Modules One through Three, logistic regressions indicated that both the SA and RRB domains made significant independent contributions to the prediction of diagnosis. However, the overall total score produced the highest predictive value, supporting the use of this score in diagnostic decisions. According to item total correlations, almost all of the items within the algorithm for Module One correlated more strongly with their assigned domain than with each other. Receiver operating characteristic curves evaluated the predictive validity of the ADOS-2 to identify individuals on the autism spectrum accurately. When comparing the ADOS to the ADOS-2, sensitivity and specificity values were largely comparable or improved with the new algorithms, particularly in Modules One and Two.

Table 3.1
ADOS-2, Module One Algorithm

Item	Code
Social Affect (SA)	
Frequency of Spontaneous Vocalizations Directed to Others	A-2
Pointing	A-7
Gestures	A-8
Unusual Eye Contact	B-1
Facial Expressions Directed to Others	B-3
Integration of Gaze and Other Behaviors During Social Overtures	B-4
Shared Enjoyment in Interaction	B-5
Showing	B-9
Spontaneous Initiation of Joint Attention	B-10
Response to Joint Attention	B-11
Quality of Social Overtures	B-12
Restricted and Repetitive Behaviors (RRB)	
Intonation of Vocalizations or Verbalizations	A-3
Stereotyped/Idiosyncratic Use of Words or Phrases	A-5
Unusual Sensory Interest in Play Material/Person	D-1
Hand and Finger and Other Complex Mannerisms	D-2
Unusually Repetitive Interests or Stereotyped Behaviors	D-4

Clinicians and Training

Clinicians included doctoral and education specialist level school psychologists and clinical psychologists, as well as one licensed social worker, all with more than two years of experience in ASD diagnostic evaluations for toddlers and preschool age children. All clinicians established research reliability on the ADOS-2, which includes achieving three consecutive co-ratings through observation with a research reliable trainer and three consecutive administrations and co-ratings with a research reliable trainer. A reliable co-rate is defined by 80% correct codes on all items in the module, as well as 80% correct codes on all items included in the algorithm.

Setting

Evaluation sessions were conducted in clinic rooms at a University clinic. Rooms were approximately 80 sq ft (8 ft by 10 ft) and included a one-way observation window that reached from ground level to the ceiling (i.e., full length). Clinic rooms contained a child size table and

chair. At least one caregiver was present with the clinician and child in the room during the entire evaluation process.

Procedures

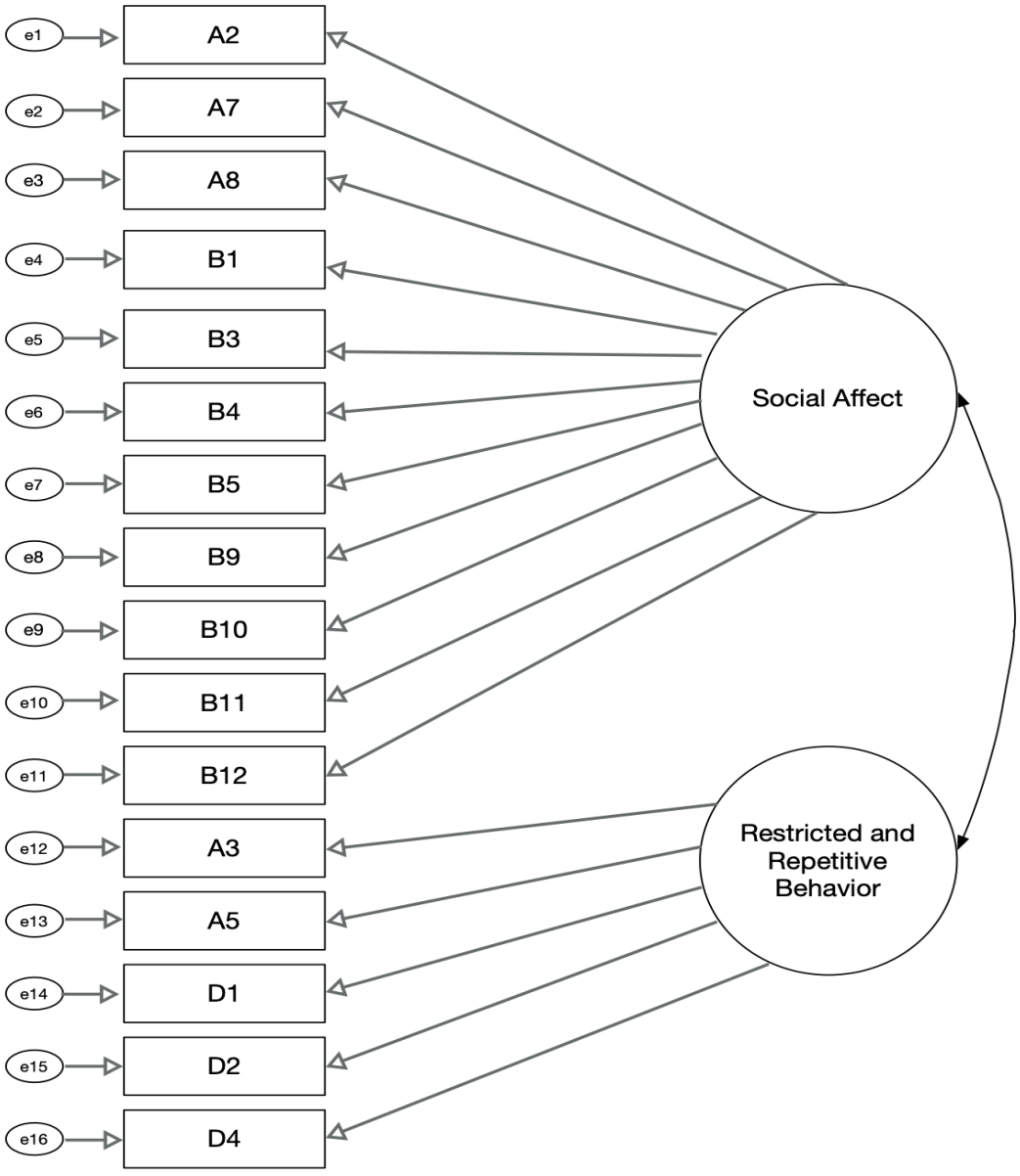
I conducted a retrospective analysis of data collected from September of 2015 through February of 2020 for participants who received the ADOS-2, Module One as part of their psychological evaluation. Participant data were deidentified and each participant was assigned a study number. Data were entered into an Excel data base and initially cleaned. Participants with missing data were excluded from the sample (N = 3 males; 1 female). Following ADOS-2 conventions, codes of 3 were converted to 2 and codes of 7, 8, or 9 were converted to 0 for items included in the ADOS-2, Module One algorithm. First, SPSS (build 1.0.0.1508) was used to compute descriptive statistics. Independent t-tests and one-way ANOVAs were conducted through SPSS to compare the mean age at evaluation and mean subscale and total scores across sex. Intercorrelations among algorithm items were also calculated for each group: the combined male and female sample; the male sample; and the female sample.

Next, the I completed a two-step psychometric evaluation of the internal structure of the ADOS-2 algorithm for each group, utilizing exploratory and confirmatory factor analyses. SPSS was used to conduct the exploratory factor analysis first through a principal component analysis (PCA) followed by common factor analysis (e.g., principal axis factoring). Data from both of these extraction methods were analyzed and recommendations comparing the methods from the literature were reviewed to determine which model was most appropriate to report and analyze. Only factors with an eigenvalue greater than one were retained (Kaiser, 1960). Further analysis of the Scree Test was conducted to confirm the suggested factors as meaningful (Cattell, 1966). A Promax oblique rotation was used based on the assumption that the factors correlate with each

other (Hendrickson & White, 1964). Loadings from Promax oblique rotations were analyzed for the combined sample and compared across the female and male samples.

Next, the proposed structure of the ADOS-2 algorithm and construct validity across groups was tested through confirmation factor analysis (CFA) using SPSS AMOS. Specifically, CFA was used to examine the fit of the two-factor model suggested by the tool and our current diagnostic criteria (e.g., DSM-5) for the combined sample, and to compare the model-fit across sex. The model as proposed by the ADOS-2 is depicted in Figure 3.1.

Figure 3.1
Structural Model Used for CFA



CHAPTER 4

RESULTS

Descriptive Statistics of Sample

Descriptive statistics for the combined, male, and female samples are reported in Tables 4.1 and 4.2. Of the 736 participants, 561 (76.2%) were male and 175 (23.8%) were female. Table 4.1 presents a breakdown of the sample by diagnosis, sex, and mean age. Overall, 89% of referred females and 86% of referred males were assigned a diagnosis of ASD. There was no significant difference in the mean age at the time of referral across sex, $t(734) = 1.281$, $p=0.201$. The ratio of males to females diagnosed with ASD from this sample was 3:1.

Table 4.1
Age at Evaluation and Diagnosis by Sex

Participant Characteristics	Combined (n=736)	Male (n=561)	Female (n=175)
Mean Age in Months	43.74	43.95	43.07
ASD	638	482	156
DD	98	79	19

As noted in previous sections, demographic data for this sample were not available; however, this information is publicly available for the school-age population within the urban setting. Although this cannot be considered a direct reflection of the sample, these data are provided as a reference. The school-age population within the urban setting report 52% of students as Black; 21% as Latinx; 14% White; 7% Asian; and 5% Multiracial/Other, with 69% of the students considered economically disadvantaged.

According to estimates of skewness and kurtosis, subscale and total scores were normally distributed for each sample. One-way ANOVAs indicated no significant difference in the Social Affect Subscales [$F(1,733) = 0.94$, $p=0.33$]; Restricted and Repetitive Behavior Subscales

[F(1,733) = 0.00, p=0.99]; and Total Scores [F(1,733) = 1.03, p=0.31] across sex. Table 4.2 presents the means and standard deviations for subscale and total scores by group.

Table 4.2
Mean, Standard Deviation, Skewness, and Kurtosis of Each Group

	Min	Max	Mean	SD	Skewness	Kurtosis
Combined Group						
Social Affect	0	20	12.21	5.40	-0.48	-0.72
RRB	0	8	4.62	2.11	-0.32	-0.63
Total Score	0	28	16.83	6.84	-0.52	-0.62
Female Group						
Social Affect	0	20	12.57	5.43	-0.61	-0.57
RRB	0	8	4.63	2.10	-0.24	-0.81
Total Score	1	28	17.31	6.69	-0.60	-0.46
Male Group						
Social Affect	0	20	12.10	5.39	-0.45	-0.75
RRB	0	8	4.62	2.12	-0.34	-0.57
Total Score	0	28	16.69	6.88	-0.50	-0.66

Correlation Analysis

Bivariate correlations were calculated across groups for all variables included within the algorithm (see Tables 4.3, 4.4, and 4.5). Correlations for most variables were significant at either the $p < .001$ or $p < 0.05$ level, with the exception of item A-5: *Stereotyped/Idiosyncratic Use of Words or Phrases*, which proved less likely to be significantly correlated with several other items. Despite its limited significance in relation to other variables, the pattern of weak correlation was consistent across the combined, male, and female groups.

The strongest correlations were identified between items included in the Social Affect Subscale. Significant but weaker correlations were identified between items included in the Restricted and Repetitive Behavior Subscale across each group. For all three groups, item A-2: *Frequency of Spontaneous Vocalization Directed to Others* and item B-4: *Integration of Gaze and Other Behaviors During Social Overtures*, represented the strongest correlation (r = 0.63 for the combined group; r = 0.62 for the male group; r = 0.66 for the female group). The female

group produced a second pair of items also correlating at r = 0.66, including item A-2: *Frequency of Spontaneous Vocalization Directed to Others* and item A-7: *Pointing*.

Table 4.3
Bivariate Correlations Between Items for the Combined Group

Item	A-2	A-7	A-8	B-1	B-3	B-4	B-5	B-9	B-10	B-11	B-12	A-3	A-5	D-1	D-2	D-4
A-2	---															
A-7	0.62**	---														
A-8	0.53**	0.53**	---													
B-1	0.50**	0.43**	0.33**	---												
B-3	0.59**	0.51**	0.50**	0.45**	---											
B-4	0.63**	0.57**	0.50**	0.52**	0.59**	---										
B-5	0.53**	0.47**	0.44**	0.37**	0.59**	0.60**	---									
B-9	0.60**	0.56**	0.47**	0.41**	0.50**	0.49**	0.40**	---								
B-10	0.48**	0.57**	0.44**	0.38**	0.51**	0.49**	0.42**	0.49**	---							
B-11	0.48**	0.50**	0.41**	0.29**	0.42**	0.47**	0.40**	0.42**	0.37**	---						
B-12	0.59**	0.48**	0.43**	0.51**	0.55**	0.56**	0.50**	0.53**	0.40**	0.36**	---					
A-3	0.29**	0.26**	0.24**	0.29**	0.26**	0.27**	0.15**	0.27**	0.17**	0.21**	0.39**	---				
A-5	-0.07*	-0.03	-0.08*	0.13**	-0.20	-0.05	-0.81*	0.00	-0.01	-0.11**	0.13**	0.24**	---			
D-1	0.38**	0.34**	0.30**	0.29**	0.30**	0.31**	0.32**	0.30**	0.24**	0.30**	0.32**	0.18**	-0.61	---		
D-2	0.29**	0.27**	0.22**	0.23**	0.22**	0.24**	0.12**	0.26**	0.15**	0.20**	0.31**	0.34**	0.14**	0.24**	---	
D-4	0.27**	0.24**	0.22**	0.24**	0.28**	0.25**	0.30**	0.23**	0.19**	0.22**	0.38**	0.20**	0.04	0.18**	0.10**	---

*significant at the $p<0.05$; **significant at the $p<0.01$

Table 4.4
Bivariate Correlations Between Items for Male Group

Item	A-2	A-7	A-8	B-1	B-3	B-4	B-5	B-9	B-10	B-11	B-12	A-3	A-5	D-1	D-2	D-4
A-2	---															
A-7	0.61**	---														
A-8	0.54**	0.56**	---													
B-1	0.48**	0.46**	0.32**	---												
B-3	0.58**	0.50**	0.50**	0.42**	---											
B-4	0.62**	0.58**	0.48**	0.52**	0.57**	---										
B-5	0.53**	0.48**	0.43**	0.36**	0.58**	0.60**	---									
B-9	0.59**	0.55**	0.48**	0.40**	0.50**	0.47**	0.40**	---								
B-10	0.46**	0.56**	0.44**	0.37**	0.50**	0.46**	0.42**	0.49**	---							
B-11	0.48**	0.51**	0.42**	0.27**	0.40**	0.47**	0.39**	0.43**	0.36**	---						
B-12	0.57**	0.49**	0.42**	0.51**	0.54**	0.56**	0.50**	0.52**	0.39**	0.35**	---					
A-3	0.29**	0.26**	0.22**	0.30**	0.22**	0.27**	0.12**	0.26**	0.14**	0.19**	0.39**	---				
A-5	-0.05	-0.03	-0.05	0.12**	-0.04	-0.03	-0.85*	-0.00	0.03	-0.11**	0.12**	0.24**	---			
D-1	0.39**	0.37**	0.33**	0.30**	0.32**	0.29**	0.33**	0.31**	0.27**	0.31**	0.33**	0.20**	-0.05	---		
D-2	0.30**	0.25**	0.24**	0.25**	0.24**	0.25**	0.10**	0.26**	0.14**	0.19**	0.33**	0.31**	0.12**	0.27**	---	
D-4	0.30**	0.24**	0.24**	0.24**	0.31**	0.27**	0.32**	0.23**	0.20**	0.23**	0.40**	0.18**	0.00	0.20**	0.09*	---

*significant at the $p<0.05$; **significant at the $p<0.01$

Table 4.5
Bivariate Correlations Between Items for Female Group

Item	A-2	A-7	A-8	B-1	B-3	B-4	B-5	B-9	B-10	B-11	B-12	A-3	A-5	D-1	D-2	D-4
A-2	---															
A-7	0.66**	---														
A-8	0.52**	0.45**	---													
B-1	0.57**	0.37**	0.35**	---												
B-3	0.63**	0.51**	0.53**	0.53**	---											
B-4	0.66**	0.55**	0.54**	0.52**	0.64**	---										
B-5	0.54**	0.44**	0.45**	0.37**	0.64**	0.58**	---									
B-9	0.65**	0.56**	0.46**	0.45**	0.48**	0.54**	0.41**	---								
B-10	0.55**	0.57**	0.47**	0.42**	0.56**	0.60**	0.44**	0.49**	---							
B-11	0.49**	0.46**	0.39**	0.34**	0.50**	0.48**	0.43**	0.41**	0.38**	---						
B-12	0.64**	0.45**	0.43**	0.53**	0.59**	0.54**	0.51**	0.57**	0.42**	0.37**	---					
A-3	0.31**	0.24**	0.31**	0.29**	0.38**	0.27**	0.24**	0.27**	0.26**	0.27**	0.38**	---				
A-5	-0.12	-0.08	-0.19**	0.17*	-0.00	-0.13	-0.84	-0.03	-0.13*	-0.12	0.18**	0.25**	---			
D-1	0.37**	0.29**	0.20**	0.28**	0.27**	0.39**	0.31**	0.25**	0.18**	0.28**	0.28**	0.14*	-0.03	---		
D-2	0.26**	0.33**	0.17*	0.16*	0.14*	0.20**	0.16*	0.28**	0.20**	0.22**	0.25**	0.42**	0.17*	0.18**	---	
D-4	0.20**	0.29**	0.16*	0.24**	0.18**	0.17*	0.23**	0.25**	0.18**	0.18**	0.35**	0.28**	0.18**	0.10	0.17*	---

*significant at the p<0.05; **significant at the p<0.01

Exploratory Factor Analysis

An exploratory factor analysis (EFA) of the internal structure of the ADOS-2 algorithm for each group was conducted first through a principal component analysis (PCA) followed by common factor analysis (e.g., principal axis factoring). Results demonstrated no difference across the two extraction methods in statistics measuring sample adequacy or fit for EFA (e.g., Kaiser-Meyer-Olkin, Bartlett's Tests of Sphericity). Furthermore, they produced the same number of factors for each group, with the same percent of variance and eigenvalues for each factor. However, slight differences were noted in the strength at which items loaded onto each factor, with the most notable changes in loadings for the female group. Further review of the literature was conducted to assist in determining which extraction method and therefore data were most appropriate to report an analyze. The literature comparing these two methods suggests: (a) common factor analysis leads to more accurate estimates of factor loadings and factor correlations for smaller numbers of underlying latent variables, (b) positive bias in estimates of loadings and corresponding negative bias in correlations among dimensions can be seen when using principal components analysis, and (c) estimates based on common factor analysis should generalize well to those obtained using confirmatory factor analytic techniques, whereas inaccurate estimates from component analysis may not be reproduced (Floyd & Widaman, 1995). In addition, previous studies exploring the factor structure of the ADOS and its various iterations have also relied on common factor analysis as an exploratory approach (Bishop et al., 2016; Gotham et al., 2007; Gotham et al., 2008). Therefore, data from the EFA using principal axis factoring with Promax rotation were analyzed and reported to verify the structure proposed by the ADOS-2 algorithm. The EFA was conducted for the combined group as well as the female and male groups individually. Kaiser-Meyer-Olkin (KMO) measures were calculated to determine the adequacy of each sample, followed by Bartlett's Tests of Sphericity, which

tested the correlation between the variables, thus indicating the appropriateness of the EFA and use of a Promax rotation. Both KMO and Bartlett's Test of Sphericity indicated the appropriateness of each sample and its suitability for analysis. See Table 4.6 for results of these tests.

Table 4.6
Tests of Sample Adequacy for Combined, Male, and Female Groups

	Combined	Male	Female
KMO*	0.94	0.94	0.91
Bartlett's Test of Sphericity: Chi-Square	4671.12** *(p< .000)*	3539.98** *(p< .000)*	1230.62** *(p< .001)*

*KMO values between 0.8 and 1 indicate the sampling is adequate
**Chi-Square test statistic is significant indicating the variables are correlated and suitable for EFA

The results of the EFAs are presented in Table 4.7 (combined), Table 4.8 (males), and Table 4.9 (females) and compare the factor solutions produced for cases in each sample. Loadings from the pattern and structure matrices are presented, along with the eigenvalues and percent variance explained for each factor. Loadings greater than 0.4 were considered adequate (Fabrigar et al., 1999).

A two-factor model, consistent with the ADOS-2 structure, was identified for the combined and male groups. The first factor can be interpreted as the Social Affect Subscale and the second as the Restricted and Repetitive Behavior Subscale. While the majority of the items fit the proposed model for the combined and male groups when interpreting the Pattern Matrix, item D-1, *Unusual Sensory Interest in Play Material/Person,* loaded onto the factor representing the Social Affect scale, which is inconsistent with the ADOS-2 algorithm. Furthermore, for the combined group, item D-2, *Hand and Finger and Other Complex Mannerisms,* and item D-4, *Unusually Repetitive Interests or Stereotyped Behaviors*, did not load strongly onto either factor. For the male group, only D-4 did not load onto either factor.

The female sample produced a different structure, indicating three factors. Similar to the combined and male groups, the first factor can be interpreted as the Social Affect Subscale. However, the second and third factors only include one item each with a loading greater than 0.4 and both items came from the Restricted and Repetitive Behavior Subscale of the algorithm (A-5 *Stereotyped/Idiosyncratic Use of Words or Phrases* and D-2, *Hand and Finger and Other Complex Mannerisms).* Therefore, three out of the five possible items included in the Restricted and Repetitive Behavior Subscale of the algorithm did not load onto any of the three factors for the female sample (A-3: *Intonation of Vocalizations or Verbalizations*, D-1: *Unusual Sensory Interest in Play Material/Person,* and D-4: *Unusually Repetitive Interests or Stereotyped Behaviors*).

Table 4.7
Factor Loadings for Combined Group

Item	Pattern Coefficients		Structure Coefficients	
	Factor 1	Factor 2	Factor 1	Factor 2
Frequency of Spontaneous Vocalizations Directed to Others (A-2)	**0.804**	0.014	**0.811**	0.402
Pointing (A-7)	**0.767**	-0.024	**0.755**	0.346
Gestures (A-8)	**0.697**	-0.072	**0.662**	0.264
Unusual Eye Contact (B-1)	**0.462**	0.266	**0.590**	0.488
Facial Expressions Directed to Others (B-3)	**0.753**	-0.008	**0.749**	0.355
Integration of Gaze and Other Behaviors in Social Overtures (B-4)	**0.784**	-0.016	**0.777**	0.363
Shared Enjoyment in Interaction (B-5)	**0.756**	-0.147	**0.685**	0.218
Showing (B-9)	**0.658**	0.073	**0.693**	0.390
Spontaneous Initiation of Joint Attention (B-10)	**0.680**	-0.082	**0.641**	0.246
Response to Joint Attention (B-11)	**0.647**	-0.102	**0.598**	0.210
Quality of Social Overtures (B-12)	**0.537**	0.360	**0.711**	0.619
Intonation of Vocalizations or Verbalizations (A-3)	0.068	**0.592**	0.354	**0.625**
Stereotyped/Idiosyncratic Use of Words or Phrases (A-5)	-0.317	**0.567**	-0.044	**0.414**
Unusual Sensory Interest in Play Material/Person (D-1)	**0.424**	0.046	**0.446**	0.250
Hand and Finger and Other Complex Mannerisms (D-2)	0.134	0.391	0.323	**0.456**
Unusually Repetitive Interests or Stereotyped Behaviors (D-4)	0.291	0.143	0.360	0.284
Eigenvalue	*6.514*	*1.458*		
% of Variance	*40.711*	*9.111*		

*Bold type indicates the strongest loading across each factor

Table 4.8
Factor Loadings for Male Group

Item	Pattern Coefficients		Structure Coefficients	
	Factor 1	Factor 2	Factor 1	Factor 2
Frequency of Spontaneous Vocalizations Directed to Others (A-2)	**0.785**	0.031	**0.800**	0.419
Pointing (A-7)	**0.783**	-0.031	**0.768**	0.357
Gestures (A-8)	**0.697**	-0.059	**0.667**	0.286
Unusual Eye Contact (B-1)	**0.445**	0.281	**0.585**	0.502
Facial Expressions Directed to Others (B-3)	**0.757**	-0.037	**0.739**	0.338
Integration of Gaze and Other Behaviors in Social Overtures (B-4)	**0.759**	0.013	**0.765**	0.389
Shared Enjoyment in Interaction (B-5)	**0.778**	-0.186	**0.686**	0.200
Showing (B-9)	**0.657**	0.065	**0.689**	0.390
Spontaneous Initiation of Joint Attention (B-10)	**0.671**	-0.089	**0.627**	0.243
Response to Joint Attention (B-11)	**0.655**	-0.121	**0.595**	0.204
Quality of Social Overtures (B-12)	**0.537**	0.352	**0.712**	0.618
Intonation of Vocalizations or Verbalizations (A-3)	0.031	**0.621**	0.338	**0.636**
Stereotyped/Idiosyncratic Use of Words or Phrases (A-5)	-0.288	**0.508**	-0.036	0.365
Unusual Sensory Interest in Play Material/Person (D-1)	**0.433**	0.072	**0.469**	0.287
Hand and Finger and Other Complex Mannerisms (D-2)	0.130	**0.400**	0.329	**0.465**
Unusually Repetitive Interests or Stereotyped Behaviors (D-4)	0.349	0.074	0.386	0.248
Eigenvalue	*6.493*	*1.444*		
% of Variance	*40.581*	*9.022*		

*Bold type indicates the strongest loading across each factor

Table 4.9
Factor Loadings for Female Group

Item	Pattern Coefficients			Structure Coefficients		
	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3
Frequency of Spontaneous Vocalizations Directed to Others (A-2)	**0.843**	-0.053	0.035	**0.844**	0.162	0.358
Pointing (A-7)	**0.645**	-0.129	0.247	**0.710**	0.077	0.477
Gestures (A-8)	**0.665**	-0.139	0.039	**0.646**	0.032	0.274
Unusual Eye Contact (B-1)	**0.605**	0.290	-0.117	**0.630**	0.417	0.177
Facial Expressions Directed to Others (B-3)	**0.836**	0.104	-0.164	**0.797**	0.279	0.185
Integration of Gaze and Other Behaviors in Social Overtures (B-4)	**0.859**	-0.087	-0.064	**0.812**	0.113	0.258
Shared Enjoyment in Interaction (B-5)	**0.727**	0.005	-0.102	**0.685**	0.155	0.183
Showing (B-9)	**0.643**	0.007	0.137	**0.699**	0.192	0.392
Spontaneous Initiation of Joint Attention (B-10)	**0.695**	-0.120	0.046	**0.683**	0.060	0.297
Response to Joint Attention (B-11)	**0.581**	-0.102	0.101	**0.596**	0.061	0.311
Quality of Social Overtures (B-12)	**0.650**	0.338	-0.022	**0.725**	0.495	0.299
Intonation of Vocalizations or Verbalizations (A-3)	0.188	0.319	0.343	0.402	0.431	**0.478**
Stereotyped/Idiosyncratic Use of Words or Phrases (A-5)	-0.306	**0.783**	0.084	-0.079	**0.723**	0.113
Unusual Sensory Interest in Play Material/Person (D-1)	0.376	-0.009	0.046	0.392	0.093	0.193
Hand and Finger and Other Complex Mannerisms (D-2)	0.023	0.113	**0.730**	0.292	0.247	**0.742**
Unusually Repetitive Interests or Stereotyped Behaviors (D-4)	0.182	0.262	0.132	0.299	0.332	0.254
Eigenvalue	*6.639*	*1.620*	*1.010*			
% of Variance	*41.492*	*10.122*	*6.313*			

*Bold type indicates the strongest loading across each factor

Confirmatory Factor Analysis

A confirmatory factor analysis (CFA) was conducted to compare the model fit across samples. The model analyzed reflected the current algorithm proposed by the tool and as presented in Figure 3.1, included 16 observed variables representing each algorithm item, which were assigned to one of two latent variables, representing the Social Affect and Restricted and Repetitive Behavior Subscales. The two latent variables were covaried and error terms included for each observed variable.

An analysis of the results of the CFA represented a similar and adequate model fit for each group, as interpreted by recommended guidelines of an RMSEA of <0.08, a CFI ≥0.90, and a RMR <0.08 (Kline, 2015). However, the TLI and GFI were below the conservative guideline of 0.95 for each group (Kline, 2015). Data demonstrating model fit for each group are presented in Table 4.10. Indices that did not meet recommended guidelines for model fit are noted in italics.

Table 4.10
Statistics for Model Fit for Combined, Male, and Female Groups

Index	Combined	Male	Female
χ^2 of Model Fit	513.723	419.915	214.021
df of Model	103	103	103
χ^2 p-value	0.000	0.000	0.000
RMSEA	0.074	0.074	0.079
CFI	0 .910	0.908	0.904
TLI	*0 .895*	*0.893*	*0.888*
RMR	0 .034	0.034	0.053
GFI	*0 .911*	*0.906*	*0.864*

Parameter estimates are presented in Tables 4.11, 4.12, and 4.13 for each group. Structural models are presented in Figures 4.1, 4.2, and 4.3. For all three groups, items included in the Social Affect subscale loaded more strongly than items included in the Restricted and Repetitive Behavior subscale. The item measuring the amount of spontaneous and purposeful

vocalizations produced by the child *(*A-2: *Spontaneous Directed Vocalizations)* demonstrated the strongest loading across groups, while the item measuring the presence of stereotyped vocalizations consistently demonstrated the weakest loading *(*A-5: *Stereotyped/Idiosyncratic Use of Words).* In addition to representing the weakest loading, item A-5 did not load significantly in the combined and male groups. Although continuing to load weakly, item A-5 was significant for the female group.

When looking specifically at the male and female groups, there was some variation in the strength of loading for items in the Restricted and Repetitive Behavior Subscale. For example, the item measuring atypical intonation patterns *(*A-3: *Intonation of Vocalizations),* loaded much more strongly onto the Repetitive and Restricted Behavior Subscale for females than it did for males. The relationship between the two subscales as latent variables noted differences across groups. For the combined and male groups, correlation estimates were 0.80 and 0.83 respectively. Weaker correlations were noted within the female group between the latent variables, indicating a correlation estimate of 0.60.

Table 4.11
Parameter Estimates for Latent Variables in the Model: Combined Group

	Unstandardized Coefficients	SE	Standardized Coefficients
Social Affect			
Spontaneous Directed Vocalizations	1.000	------	0.811
Pointing	0.972*	0.043	0.749
Gestures	0.828*	0.044	0.655
Unusual Eye Contact	0.745*	0.043	0.609
Facial Expressions Directed to Others	0.762*	0.034	0.746
Integration During Social Overtures	0.836*	0.035	0.777
Shared Enjoyment in Interaction	0.857*	0.043	0.676
Showing	0.901*	0.044	0.700
Spontaneous Initiation of Joint Attention	0.910*	0.049	0.640
Response to Joint Attention	0.788*	0.048	0.579
Quality of Social Overtures	0.753*	0.035	0.720
Restricted and Repetitive Behaviors			
Intonation of Vocalizations	1.000	------	0.483
Stereotyped/Idiosyncratic Use of Words	0.126	0.088	0.063
Sensory Interest in Play Material/Person	1.077*	0.121	0.512
Hand/Other Complex Mannerisms	0.945*	0.115	0.448
Repetitive Interests/Stereotyped Behaviors	0.723*	0.093	0.406

*Indicates significant at the $p<0.05$

Table 4.12
Parameter Estimates for Latent Variables in the Model: Male Group

	Unstandardized Coefficients	SE	Standardized Coefficients
Social Affect			
Spontaneous Directed Vocalizations	1.000	----	0.801
Pointing	1.008*	0.051	0.761
Gestures	0.857*	0.052	0.663
Unusual Eye Contact	0.741*	0.050	0.602
Facial Expressions Directed to Others	0.753*	0.040	0.734
Integration During Social Overtures	0.836*	0.042	0.769
Shared Enjoyment in Interaction	0.863*	0.051	0.675
Showing	0.915*	0.052	0.695
Spontaneous Initiation of Joint Attention	0.911*	0.059	0.627
Response to Joint Attention	0.778*	0.055	0.575
Quality of Social Overtures	0.755*	0.041	0.717
Restricted and Repetitive Behaviors			
Intonation of Vocalizations	1.000	-----	0.445
Stereotyped/Idiosyncratic Use of Words	0.085	0.104	0.040
Sensory Interest in Play Material/Person	1.241*	0.160	0.541
Hand/Other Complex Mannerisms	0.997*	0.145	0.432
Repetitive Interests/Stereotyped Behaviors	0.802*	0.119	0.420

*Indicates significant at the $p<0.05$

Table 4.13
Parameter Estimates for Latent Variables in the Model: Female Group

	Unstandardized Coefficients	SE	Standardized Coefficients
Social Affect			
Spontaneous Directed Vocalizations	1.000	----	0.846
Pointing	0.846*	0.078	0.715
Gestures	0.745*	0.080	0.640
Unusual Eye Contact	0.765*	0.083	0.636
Facial Expressions Directed to Others	0.779*	0.064	0.783
Integration During Social Overtures	0.826*	0.065	0.798
Shared Enjoyment in Interaction	0.831*	0.083	0.676
Showing	0.845*	0.079	0.713
Spontaneous Initiation of Joint Attention	0.902*	0.089	0.685
Response to Joint Attention	0.809*	0.096	0.590
Quality of Social Overtures	0.742*	0.067	0.727
Restricted and Repetitive Behaviors			
Intonation of Vocalizations	1.000	-----	0.700
Stereotyped/Idiosyncratic Use of Words	0.317*	0.134	0.221
Sensory Interest in Play Material/Person	0.464*	0.137	0.328
Hand/Other Complex Mannerisms	0.786*	0.156	0.556
Repetitive Interests/Stereotyped Behaviors	0.504*	0.125	0.405

*Indicates significant at the $p<0.05$

Figure 4.1
Structural Model with Standardized Coefficients: Combined Group

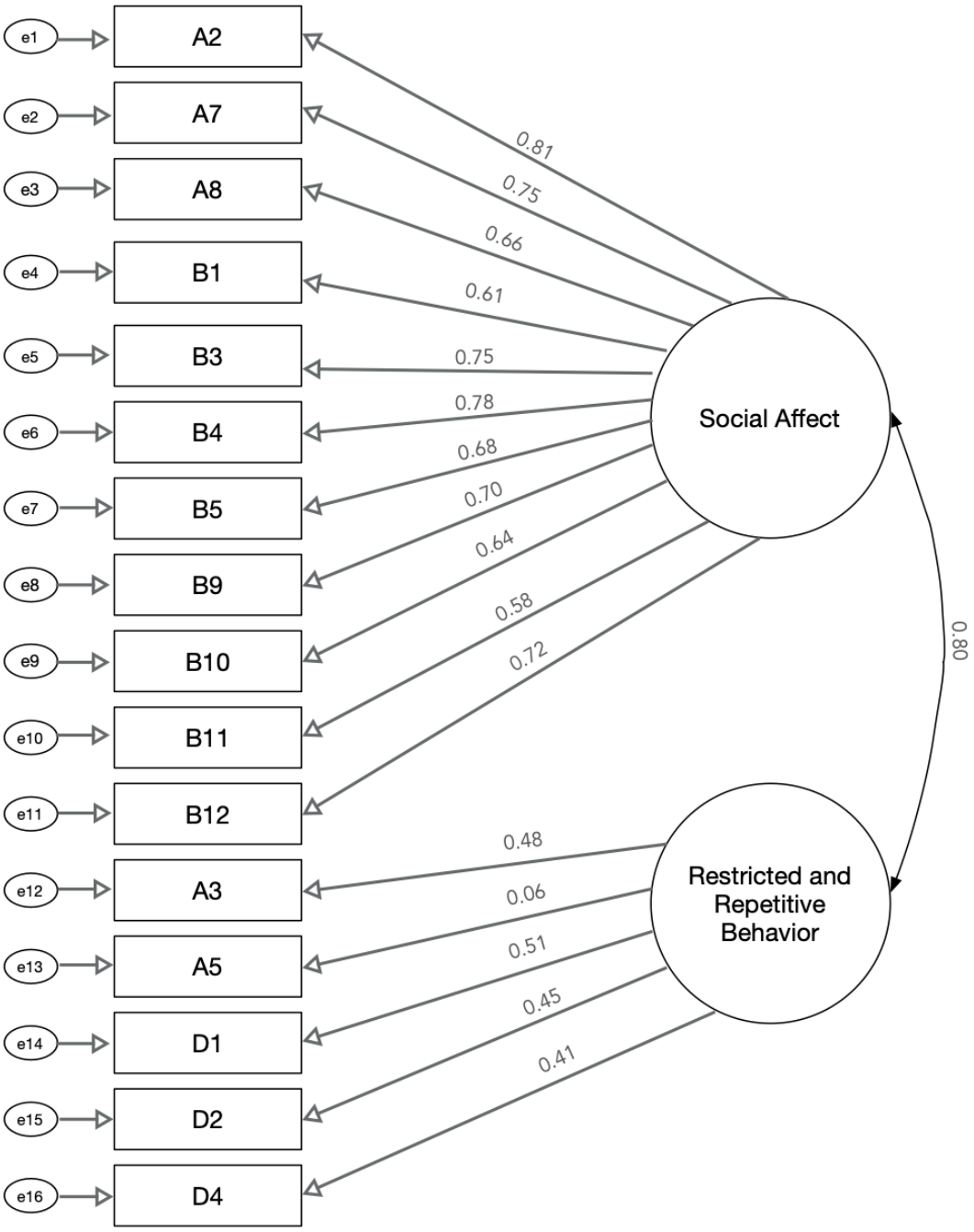


Figure 4.2
Structural Model with Standardized Coefficients: Male Group

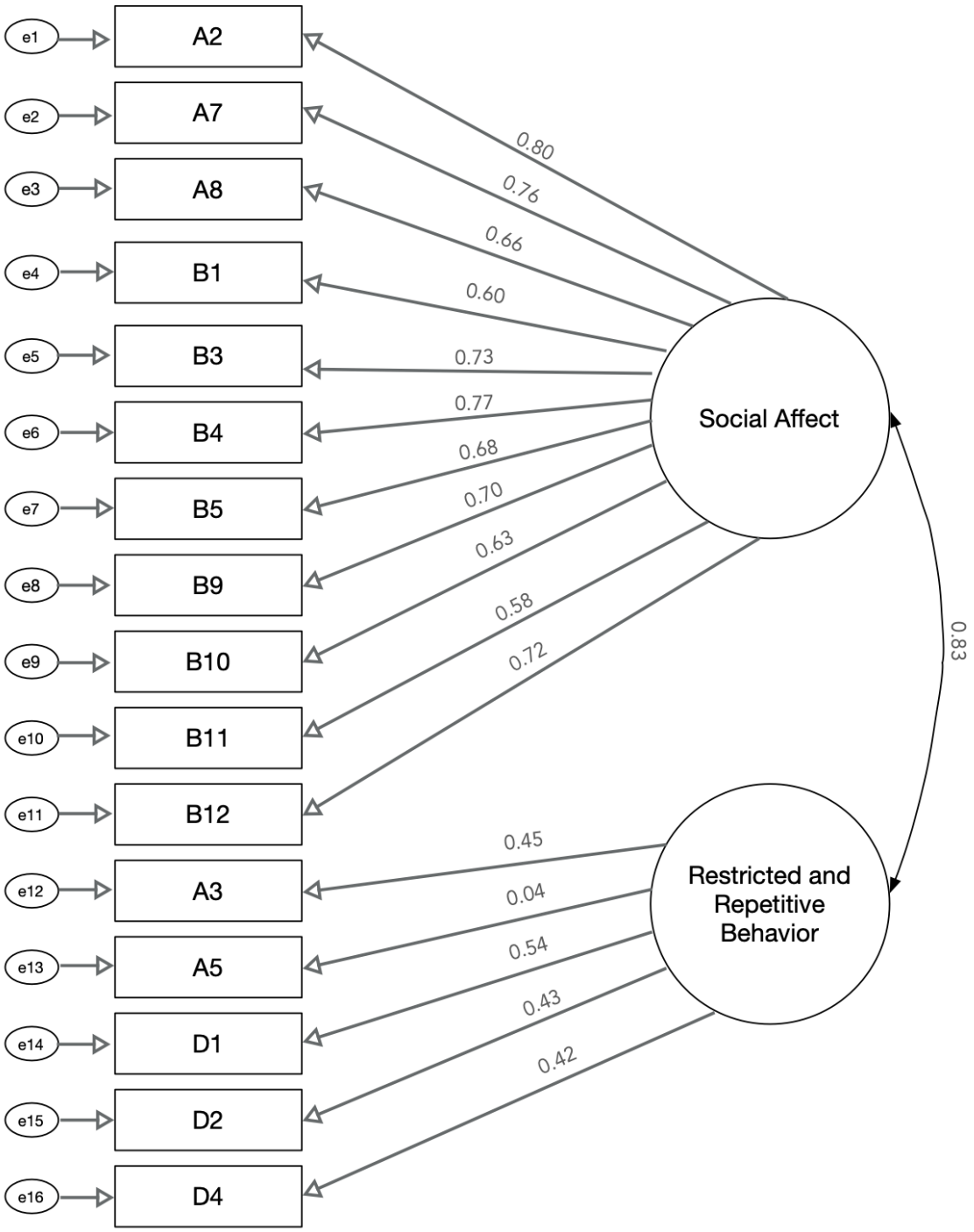
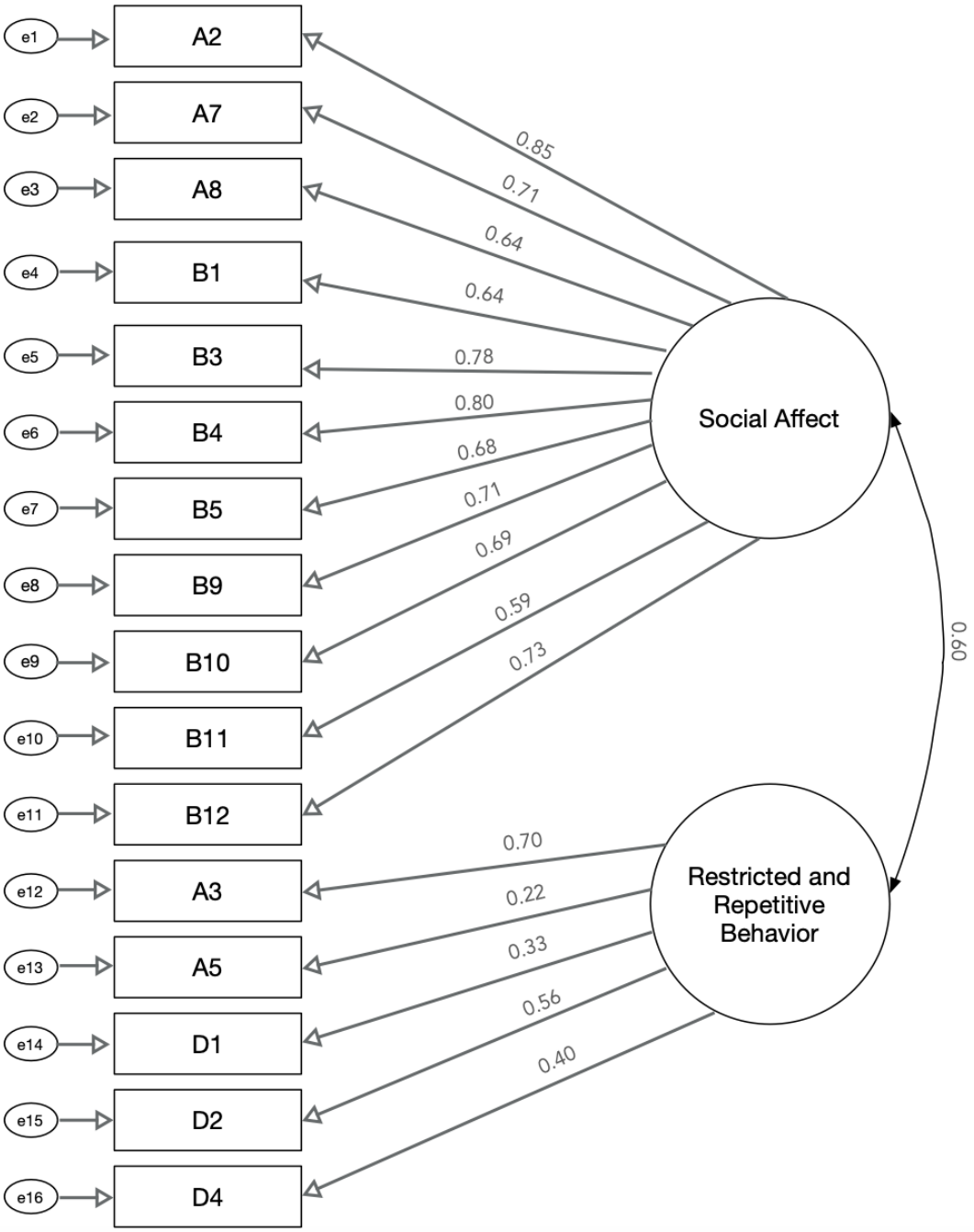


Figure 4.3
Structural Model with Standardized Coefficients: Female Group



When looking specifically at item loading strengths across the male and female groups, the items representing directed vocalizations and the integration of communicative strategies represent the strongest loadings. When considering the subsequent items in rank order for the male group, the rest of the items included in the Social Affect scale are noted: pointing, directed facial expressions, the quality of social overtures, showing, sharing enjoyment, gestures, initiating joint attention, the use of eye contact, and the child's response to joint attention. For the male group, the items included in the Restricted and Repetitive Behavior Subscale follow. The female group produced a slightly different rank order, which included one item (A-3: *Intonation of Vocalizations*) from the Restricted and Repetitive Behavior Subscale loading more strongly than some items from the Social Affect Subscale. Table 4.14 presents a clearer comparison of item loadings across the male and female groups in rank order.

Table 4.14
Order of Item Loadings for Male and Female Groups

Rank	Male Group	Female Group
1	Spontaneous Directed Vocalizations (SA)	Spontaneous Directed Vocalizations (SA)
2	Integration During Social Overtures (SA)	Integration During Social Overtures (SA)
3	Pointing (SA)	Facial Expressions Directed to Others (SA)
4	Facial Expressions Directed to Others (SA)	Quality of Social Overtures (SA)
5	Quality of Social Overtures (SA)	Pointing (SA)
6	Showing (SA)	Showing (SA)
7	Shared Enjoyment in Interaction (SA)	Intonation of Vocalizations (RRB)
8	Gestures (SA)	Spontaneous Initiation of Joint Attention (SA)
9	Spontaneous Initiation of Joint Attention (SA)	Shared Enjoyment in Interaction (SA)
10	Unusual Eye Contact (SA)	Gestures (SA)
11	Response to Joint Attention (SA)	Unusual Eye Contact (SA)
12	Sensory Interest in Play Material/Person (RRB)	Response to Joint Attention (SA)
13	Intonation of Vocalizations (RRB)	Hand/Other Complex Mannerisms (RRB)
14	Hand/Other Complex Mannerisms (RRB)	Repetitive Interests/Stereotyped Behaviors (RRB)
15	Repetitive Interests/Stereotyped Behaviors (RRB)	Sensory Interest in Play Material/Person (RRB)
16	Stereotyped/Idiosyncratic Use of Words (RRB)	Stereotyped/Idiosyncratic Use of Words (RRB)

Note: SA=Social Affect Subscale; RRB=Restricted and Repetitive Behavior Subscale

CHAPTER 5

DISCUSSION

While many questions remain unanswered about autism spectrum disorder (ASD), decades of research and clinical practice have revealed a greater understanding of the disorder's etiology, symptom presentation, and prevalence. There is a general consensus in the scientific community that indicates the likelihood of an interaction between one's genes, biology, and the environment that make a child more likely to have an ASD, along with risk factors, such as low birth weight, advanced parental age, and the presence of birth complications, among others (Modabbernia et al., 2017). In addition, the diagnostic criteria have evolved significantly over time, from the initial conceptualization of autism as a childhood presentation of schizophrenia, to a separate diagnosis which emphasizes a chronic course of social and behavioral differences emerging in the infant and toddler years. Lastly, researchers worldwide have identified a concerning increase in prevalence of the disorder in the general population. Most recent estimates by the Centers for Disease Control (CDC) indicate that 1 in 54 children in 2016 met criteria for ASD in the United States, with rates nearly equal among all racial, ethnic, and socioeconomic groups. Still, studies continue to demonstrate consistently disproportionate rates of ASD in males to females (Maenner et al., 2020).

Researchers and clinicians have explored possible explanations for differences in the sex ratio, including suggestions about the nature and etiology of ASD, such as the extreme male brain (Baron-Cohen, 2002), the female protective genetic effect (Jacquemont et al., 2014), the ability of females to "camouflage" symptoms (Dean et al., 2017), and the presence of a separate autism phenotype for females (Hull et al., 2020; Kopp & Gillberg, 1992). Others have suggested

bias in the assessment and diagnostic processes which could influence the likelihood that males receive a diagnosis more often and earlier than females (Loomes et al., 2017).

The issue of bias in diagnostic criteria and psychological assessment is hardly a new concept. A starting point for this research includes critically analyzing the tools with which we measure the presentation of symptoms. Previous research examining bias in the predictive and construct validity of assessment tools has relied on methods like partial correlation procedures, chi-square techniques, path analysis, and factor analysis (review in Reynolds & Suzuki, 2012). In particular, factor analysis allows researchers to examine the structure of a tool and compare results for two or more groups, such as sex, to determine if the test functions similarly from one group to another.

The Autism Diagnostic Observation Schedule (ADOS) (Lord et al., 1999) and its subsequent revision are considered the gold-standard in the assessment of autism symptoms and sometimes used exclusively as gatekeepers for a clinical diagnosis, eligibility for benefits through insurance agencies, or eligibility for participation in research. Studies that established and replicated the psychometric properties of this tool and others measuring autism symptoms have used samples that were largely disproportionate in the male to female ratio (Gotham et al., 2007; Gotham et al., 2008; Lord et al., 1999; Lord et al., 2012). This has led researchers to look more critically at this and other tools to question whether they accurately capture symptom severity in females. Studies have compared the total scores generated by the ADOS across sex, with mixed findings. Others have explored the factor structure and model fit of the tool across samples which differ in language level and diagnosis (e.g., with ASD or without ASD), though few to no studies have been published that analyze the tool's factor structure and construct validity across sex. Furthermore, the sample used to validate and replicate the tool included

primarily male, white, and American individuals who met criteria for ASD. Critics have expressed concern that this may limit the generalizability of the tool's results to a broader population due to such factors as sex or culturally appropriate social behaviors (e.g., eye contact) (McCrimmon & Rostad, 2014).

Findings from Descriptive Analyses

To address existing gaps in the literature, the current study examined the structure of the ADOS-2 across sex through exploratory and confirmatory factor analysis in a sample of 736 preschool age children. The736 cases who received an assessment with the ADOS-2, Module One, were extracted from a larger sample which included administrations of other Modules, requiring stronger language skills. While data were not available to formally identify other characteristics to infer a cohesive sample, such as standard scores of cognitive functioning or language development, the appropriateness of administering a Module One for these children indicates that all participants shared similar verbal language abilities. That is, their verbal language ranged from the use of no words to one-to-two-word utterances, consistent with the ADOS-2 guidelines for a fit with the administration of Module One (Lord et al., 2012).

This particular sample represented a clinical, at-risk population of children referred through the early intervention system following screening and broad developmental assessment in the community. All children presented with developmental delays and were demonstrating valid concern for ASD symptoms as evidenced by high rates of diagnosis following the evaluation, including 89% of referred females and 86% of referred males. Similar to other studies relying on screened or high-risk samples (Loomes et al., 2017), the ratio of males to females with ASD was lower than measured in the general population (i.e., 4:1) with 3 males to every 1 female meeting criteria in this study. In a metanalysis of sex ratio across studies, Loomes

and colleagues (2017) noted lower ratios of males to females, ranging from 2:1 to 3.5:1, in designs which actively sought cases through population-based screening and assessment, similar to the child-find procedures imbedded within the early intervention system and as recommended best practice for pediatric primary care. That is, the sample represented a group of cases where previous concern for development had been identified through standard screening and surveillance procedures by pediatricians and/or daycare and preschool providers, as well as concern expressed by caregivers. Further, this sample was subsequently screened by a clinical evaluation team, where differential "red flags" for autism had been suggested, prompting a referral for additional assessment. Therefore, the ratio found in this study may provide additional evidence that differences in autism sex ratios are indeed less when early delays or symptoms are screened and surveilled in the general population and confirmed through comprehensive and careful assessment.

On the other hand, these results could also suggest that the high-risk sample included only the more significantly impaired cases, or cases where autism symptoms were clearer. This would be consistent with studies that have found that females who present with a greater number of behavioral difficulties or cognitive deficits are referred at similar rates as males, while females with mild symptoms are significantly under-referred (Duvekot et al., 2017; Dworzynski et al., 2012; Shattuck et al., 2009). The latter is further supported by an analysis, using one-way ANOVAs, of the total and subscale scores produced for all participants, which did not indicate significant differences across sex. Again, this is consistent with previous analyses where no sex differences in ADOS total scores were found in similar high-risk samples (Duvekot et al., 2017). In contrast, studies that identified significant differences in total ADOS scores across sex included samples that were considered both high and low-risk (Zwaigenbaum et al., 2012). This

suggests that samples of individuals with presumably mild or no symptoms of autism demonstrate differences in the tool's ability to measure key social, communicative, and behavioral skills across sex. Overall, descriptions of sex ratio and consistency in total scores from this sample suggest that the true male to female ratio may be lower in the general population when symptom presentation is first identified through population-based screening procedures. However, even our sample may have included cases with more obvious "red flags", resulting in little difference in total scores or symptom severity across sex. Therefore, it remains possible that more females with less-detectable "red flags" go unscreened or are not captured by broad screening and surveillance procedures in the community. If this is the case, then the sex ratio could be even less disparate than indicated in this sample or in other samples that explore high risk cases as identified through standard screening procedures.

Another unique aspect of this sample included an average age of diagnosis that was lower than the national average of 52 months (Maenner et al., 2020). The CDC's most recent review of prevalence indicates that 85% of children identified with ASD had concerns about their development noted in their records by 36 months of age (Centers for Disease Control, 2020). However, there is frequently a lag between first concern and first developmental evaluation, which may affect when children with ASD can begin to access the services they need. Racial and ethnic differences still exist in how early ASD is identified through evaluation and diagnosis. Overall, Black and Latinx children with ASD receive evaluations later than White children (Maenner et al., 2020). The mean age at the time of referral in this sample for males and females was 43.7 months, eight months younger than the national average. While demographic data for the current study's sample were not available, the racial and ethnic breakdown of school-age children in the same urban center includes a majority minority population, with 73% Black and

Latinx students as compared to 14% White students. This may suggest that minority children in this sample received a diagnosis and access to services sooner than the national average, and even earlier still as compared to minority groups in US samples (Maenner et al., 2020). Furthermore, when using school-age demographic data as a reference group, this sample demonstrates a significantly more diverse racial group than samples used to validate and replicate the ADOS-2, which ranged from 71%-91% white participants (Lord et al., 2012).

The analysis of descriptive data from this study is an important addition to the literature describing sex ratio, symptom severity, and mean age of diagnosis in a clinical, high-risk sample which uniquely included a large number of preschool age children from an urban minority population. While evaluation of potential bias in the screening and referral process is beyond the scope of this study, the high and nearly equal rate of positive ASD diagnoses across sex in this sample may suggest lack of bias in the diagnostic process itself. That is, of children referred in this sample, nearly equal rates of females and males received a diagnosis of ASD based on the DSM-5 criteria. This further supports the appropriateness and adequacy of the sample for isolating potential differences across sex in the primary tool used to assess symptoms.

Findings from Factor Analyses

The first aim of this study was to explore the ADOS-2, Module One algorithm for the combined group and across sex to examine the underlying factor structure and correlations among variables. This aim intended specifically to determine if the male and female groups yielded a two-factor model representing the subscales of the tool and if items within and across each subscale were correlated at similar levels across sex. Correlations among all variables were significant, with the exception of one item A-5, *Stereotyped / Idiosyncratic Use of Words or Phrases*; a pattern consistently observed across the combined, male, and female groups. Further,

item A-5 was the only item negatively correlated with other items across each group. In other words, when more significant social, communicative, and behavioral symptoms were observed, less scripted and idiosyncratic speech was noted. This pattern is not unexpected since this item may not be included in the algorithm of a child who presents with no verbal language. Furthermore, the pattern is consistent with clinical expectations of very young children with more significant social and communication impairments who have less overall verbal speech and therefore less language to categorize as stereotyped or idiosyncratic. In turn, as children with significant social and communication symptoms acquire verbal speech over time, it is not uncommon for that speech to initially present as highly repetitive and stereotyped. Studies that have previously explored the factor structure of other modules of the ADOS-2, appropriate for older children with more verbal language, found item A-5 to load strongly onto the Repetitive and Restricted Behavior Subscale. While these studies did not explore structural differences across sex, this information may have clinical implications for diagnosticians as they use different modules across age and language groups (Bishop et al., 2016).

Although most correlations were significant, relationships between variables did not exceed a value of 0.66 and were therefore only moderately correlated (Goodwin & Leech, 2006). As expected, still moderate but stronger positive correlations were identified between items included in the Social Affect Subscale, consistent with studies that originally validated the algorithm (Gotham et al., 2007; Gotham et al., 2008; Lord et al., 2012). Items in this scale represent symptoms falling within the social communication domain of the DSM-5 diagnostic criteria. Individuals who meet criteria for a diagnosis are required to demonstrate three out of three social communication symptoms, whereas only two out of four symptoms in the repetitive and restricted behavior domain are required (APA, 2013). The monothetic approach to social

communication symptoms in ASD criteria is well-understood and accepted as a reflection of the fact that foundational social difficulties underlie ASD and differentiate its presentation from other neurodevelopmental disorders (APA, 2013). Therefore, that individual items in this domain are highly correlated is an expected result. As such, significant but weaker correlations were identified between items included in the Restricted and Repetitive Behavior Subscale across each group. This is a similar finding to validation and replication studies of the ADOS-2 algorithm, which found items included in the Restricted and Repetitive Behavior domain to demonstrate lower internal consistency and weaker correlations overall (Lord et al., 2012; McCrimmon & Rostad, 2014).

Specific relationships between social symptoms differed slightly across male and female groups. Both groups demonstrated the strongest relationship between social symptoms measuring the frequency and quality of purposeful communication (A-2 and B-4). However, only the female group produced another equally strong relationship between frequently used vocalizations and pointing (A-2 and A-7). This indicates that females in this sample who utilized this critical developmental gesture more often and more effectively, also used more purposeful vocalizations when communicating. In summary, for both males and females in this sample, items included in the ADOS-2, Module One algorithm consistently reflected theories behind the DSM-5 criteria, noting strong and positive relationships in social symptoms and significant but perhaps less meaningful relationships between behavioral symptoms.

Exploratory Factor Analysis: Combined and Male Groups

Exploratory factor analysis (EFA) compared the factor solutions produced for cases in each group resulting in a two-factor model, consistent with the ADOS-2 structure, for the combined and male groups. As expected, the two factors represented the Social Affect Subscale

and Restricted and Repetitive Behavior Subscale of the ADOS-2 and were consistent with the two domain criteria suggested by the DSM-5 (APA, 2013). Considering more males overall were represented in the sample, it is not surprising that results from EFAs with the combined and male groups were similar. Interestingly, results from this sample were not consistent with findings from Gotham and colleagues (2007), which indicated a third factor of Joint Attention, including items measuring gestures, eye contact, joint attention, and showing.

While most item loadings reflected the suggested pattern of the ADOS-2 subscales and the DSM-5 criteria, the item measuring sensory seeking behavior (D-1) inaccurately loaded onto the factor representing the Social Affect Subscale and the item measuring repetitive and stereotyped behaviors and sensory sensitivity (D-4) did not load strongly onto either factor. The fact that items measuring sensory differences were more strongly related to social symptoms (D-1) or not strongly related to either factor (D-4) is noteworthy, as sensory features of autism represent the "newest" symptom to be added to the current criteria. As part of the DSM-IV field trial, the potential criterion of hypo- or hypersensitivity to sensory stimuli was examined but not included in the final definition, since it was a less powerful diagnostic feature than other potential criteria. In contrast, the DSM-5 definition of autism includes sensory issues as one of the four restricted and repetitive behavior features defined as "hyper or hypo reactivity to sensory input or unusual interest in sensory aspects of the environment" (Grapel et al., 2015). Sensory issues have long been an associated autism feature, and specifically related to characteristics of restricted and repetitive behavior, thus prompting its inclusion as a symptom within that specific domain. Previous studies have demonstrated that high levels of sensory sensitivity predict high levels of repetitive behaviors and, similarly, increased sensory seeking behaviors are highly associated with ritualistic-like behavior in children with ASD (Boyd et al., 2010; Schultz &

Stevenson, 2019). This line of research did not find a strong relationship between sensory behavior and social communication deficits in autism, across tools relying on both direct observation and caregiver report (Boyd et al., 2010; Schultz & Stevenson, 2019).

The weak relationship between either factor and item D-4 in the combined and male group as it relates to both sensory and behavioral features is also worthy of further discussion. In addition to measuring hyper-responsiveness to sensory stimuli, such as auditory or tactile sensitivity, item D-4 measures the presence of repetitive, ritualistic, or stereotyped behavior and interests, making it an item that includes aspects of four out of four symptoms in the restricted and repetitive behavior domain of the DSM-5 criteria (APA, 2013; Lord et al., 2012). For this reason, it may be predicted that item D-4 would be most representative of autism behavioral symptomology and therefore maintain the strongest relationship to the Restricted and Repetitive Behavior Subscale of the ADOS-2 algorithm. Considering our data demonstrated the contrary, it is possible that the item as a "catch all" for behavioral symptoms in this age group or population specifically, weakens its relationship or effectiveness at predicting symptoms in this subscale. This may also indicate that a singular tool like the ADOS-2 is more accurate at measuring social communication symptoms than restricted and repetitive behavior, a theory which prompted the developers to originally exclude these behavioral items from the algorithm (Lord et al., 1999; McCrimmon & Rostad, 2014). Evidence from the validation and replication studies also noted that when added to the algorithm, lower test-retest values were achieved for items within the Restricted and Repetitive Behavior Subscale (Lord et al., 2012). While concern for the context of the ADOS as too time-limited to capture behavioral symptoms was questioned and the inclusion of these items into the algorithm produced improved results for most Modules (Gotham et al., 2007; Lord et al, 2006), there appears to be room for ongoing research into the contribution of

Restricted and Repetitive Behavior Subscale items to the validity and reliability of the tool across populations.

Exploratory Factor Analysis: Female Group

As predicted, the female group produced a different structure than the combined and male groups, indicating three factors, inconsistent with the ADOS-2 algorithm and the DSM-5 diagnostic criteria. Similar to the combined and male groups, the first factor seems to represent the Social Affect Subscale of the ADOS-2. However, the second and third factors only include one item each with a loading greater than 0.4, both from the Restricted and Repetitive Behavior Subscale of the algorithm. These two items measured the presence of stereotyped or idiosyncratic language (A-5) and atypical and repetitive motor mannerisms (D-2). The three other items included in the Restricted and Repetitive Behavior Subscale of the algorithm did not load onto any of the three factors for the female group (i.e., A-3, D-1, and D-4). While previous research (Gotham et al., 2007) also identified a three factor model for Module One in combined groups of both male and females, their factor breakdown included two factors comprised of a combination of social and communication symptoms and a third including symptoms from the restricted and repetitive behavior domain. Therefore, it is interesting to note the different pattern of loadings produced when female-only groups are analyzed within the same Module of the ADOS.

Considering the results from the EFA, the Restricted and Repetitive Behavior Subscale of the ADOS-2 does not appear well-constructed to represent behavioral symptoms for either group, though even less so for preschool age females. Previous studies have implicated the behavioral symptom domain as less predictive of a diagnosis for females as measured by standardized tools, though not the ADOS-2 specifically. These studies found males with ASD to

exhibit higher levels of repetitive and stereotyped behaviors than females (Beggiato et al., 2017; Hartley & Sikora, 2009; Szatmari et al., 2012), with less to no significant sex differences in social interaction or communication (Van Wijngaarden-Cremers et al., 2014). The poorly constructed factor representing the Restricted and Repetitive Behavior Subscale of ADOS-2 for females adds support to this literature, particularly within a young sample of children.

Confirmatory Factor Analysis

A confirmatory factor analysis (CFA) compared the model fit across groups. The model reflecting the current algorithm proposed by the tool was used for each group, with the Social Affect and Restricted and Repetitive Behavior Subscales representing two latent variables. The model represented a good fit for each group across the majority of fit indicators (Kline, 2015). Although the EFA demonstrated differences in factor structure between the male and female groups, consistent with our hypothesis, when two factors were forced, the fit proved appropriate for the combined group and across sex, which is a surprising finding.

An analysis of the standardized coefficients produced through the CFA, across groups, demonstrated similar results to the EFA, where items included in the Social Affect Subscale loaded more strongly than items included in the Restricted and Repetitive Behavior Subscale. Again, the item measuring the amount of spontaneous and purposeful vocalizations produced by the child (A-2) represented the strongest loading for the combined, male, and female groups. The item measuring the presence of stereotyped vocalizations consistently demonstrated the weakest loading *(*A-5). In addition to representing the weakest loading, item A-5 did not load significantly in the combined and male groups, though it remained significant for the female group. As observed in the EFA, overall, items included in the Restricted and Repetitive Behavior Subscale produced weaker loadings for the female group, when compared to the male

group, with the exception of item A-3, which measures the presence of atypical intonation patterns and voice quality. This item loaded at 0.70 for the female group as compared to 0.45 for the male group. The results of the CFA support evidenced gleaned from the EFA, which indicates a poorly constructed Restricted and Repetitive Behavior Subscale for females as compared to males, both when factor structure is explored and when the algorithm as suggested by the tool is applied.

When looking specifically at the strength of item loadings across the male and female groups, items representing directed vocalizations and the integration of communicative strategies represent the strongest loadings. Overall, social symptoms remain the stronger predictors of an ASD diagnosis for both males and females in this sample, though there are minor differences in loadings even within this scale across sex.

Previous studies have explored the key items from the ADOS-2, Module One, which predict ASD diagnosis across the early developmental period. Chawarska and colleagues (2014) looked at specific items included in Module One in a group of high-risk siblings (i.e., younger siblings of children diagnosed with ASD) at 18 and 36 months. Although they did not explore data across sex, they noted differences in items that predicted ASD diagnosis at each time point. For example, at 18 months, the item measuring eye contact (B-1) by itself was a poor prognostic indicator, though minimal eye contact that co-occurred with the limited use of gestures (A-8) increased the likelihood significantly of the child meeting criteria for ASD by 36 months. Further, 18 month old children with strong eye contact (B-1) but emerging repetitive and restricted behaviors (D-4) were much more likely to meet criteria for ASD at 36 months. When items were explored overall, eye contact (B-1), the presence of repetitive and restricted behaviors (D-4), the use of gestures (A-8), and atypical intonation patterns (A-3) were a few of the

strongest predictors of ASD diagnosis at 36 months, along with other items from the tool not included in the algorithm (e.g., the tendency to give objects within play, the presence of functional or creative play) (Chawarska et al., 2014). While the authors did not share data regarding which items predicted an ASD diagnosis at or after 36 months, they noted an increase from 18 to 36 months in the overall severity of symptoms, as measured by the ADOS-2, indicating this time period to be critical in the decline or plateau of social communication development and increase in behavioral symptoms.

The work of Chawarska and colleagues (2014) provides insight into the importance of recognizing individual or groups of items that more strongly predict an ASD diagnosis for specific groups, which in their study included cases within a particular developmental period. Given the developmental nature of ASD and the importance of identifying risk early on, analyzing symptom features as unique to specific time points that may not generalize to other developmental periods seems critical. For example, the identification of combinations of features specific to 6-, 12-, 18-, or 24-month age levels could contribute to an understanding of the process that underlies the emergence of social deficits and behavioral atypicality. While this analysis did not explore factor structure or item level correlation across developmental period, it instead analyzed item level predictors across sex. Therefore, it is possible to discuss similarities and differences in specific items or groups of items that contribute to an increased ADOS-2 score and likely diagnosis of ASD for groups of males and females at preschool age. As noted previously, the frequency and quality of directed and purposeful vocalizations (A-2, B-4) most strongly predicted ASD in males and females. Next, items measuring pointing, the use of facial expressions, the overall quality of social overtures, and the tendency to show toys or objects as a way of sharing interests were the strongest predictors for males and females. After this group of

items, males and females differed slightly in that the presence of atypical voice quality or intonation patterns demonstrated the next strongest loading in females, over other items that measure classic social communication symptoms, including joint attention and the use of gestures. In contrast to the findings of Chawarska and colleagues (2014) in 18 month old children, eye contact (B-1) and repetitive and restricted behavior (D-4) were 10th and 15th in rank for males and 11th and 14th for females when all 16 items were ordered according to the strength of factor loadings. When used with this preschool age sample, minor differences are noted in ADOS-2 items as strong prognostic indicators across sex and more significant differences seem apparent when compared to a younger cohort of children included in Chawarska and colleague's high risk sibling sample.

The current study addressed a clear gap in the literature, exploring the utility of a "gold standard" assessment tool across sex for a uniquely diverse, community-based sample. Consistent with previous studies, overall findings demonstrate the tool's strength in identifying social-communication symptoms, though questions remain with its ability to capture important behavioral symptoms both within and across sex groups. Differences in factor structure across sex resulted from exploratory analyses, though the application of the model suggested by the tool indicated good model fit across sex. Perhaps the most interesting results lie in the review of specific items and item groups most indicative of ASD diagnosis across sex for this specific age group. Furthermore, results suggest important implications for the tool's use in clinical practice as well as room for further study of Module One through population-based (low-risk) samples as well as other Modules across sex.

Implications for Practice and Future Directions of Study

The current study presents many important implications for both research and clinical practice. First, this discussion reemphasizes suggestions made by Ratto and colleagues (2018), which caution the limited generalizability of results from studies that use specific measures, including the ADOS-2, as gate-keepers for participant eligibility. While best practice within psychological assessment cautions diagnostic conclusions made from a singular tool, the reality within community clinical practices is that specific scores from tools like the ADOS-2 determine diagnosis and therefore eligibility for services and sometimes insurance benefits. Further, it is difficult for the field to move forward in improving this or other tools when research uses this same tool as a singular criterion for participation and relies on samples that are primarily limited to white males. Continuing discussions about sampling bias, diagnostic bias, and biases resulting from homogenous validation samples should be ongoing. As clinicians become part of conversations about policy, including access to services, this knowledge and this evidence base are critical.

Data from this study and other studies demonstrating a less disparate sex ratio dependent upon sampling procedures and high versus low-risk populations lead to many questions regarding how well the community-based surveillance and screening procedures identify females with higher cognitive ability though with mild ASD symptoms, at least as captured by current screening and assessment tools. Future research and clinical practice should continue to explore screening and surveillance practices across sex, for example, examining the ability of widely used screeners to identify symptom "red flags" in females. Furthermore, as tools like the ADOS-2 continue to be analyzed and hypotheses about a female autism phenotype explored, clinicians

and researchers may be motivated to adopt sex-based screeners or norms that are more sensitive to early "red flags" across sex.

Ongoing questions about the ability of the ADOS-2 to capture symptoms of repetitive and restricted behavior were further analyzed in this study, with results suggesting a poorly constructed domain within and across groups. Indeed, the historical context of the development of the ADOS algorithms, namely the initial exclusion of a Restricted and Repetitive Behavior Subscale, is not information that is introduced to clinicians in the field during their training and requires a deep-dive into the literature to become apparent. However, this information seems most critical to practitioners within the field using the ADOS-2 Module One, who may find children with sub-clinical restricted and repetitive behavior symptoms as measured within the context of the assessment, leading to an algorithm score below the cut-off. Clinicians who rely strongly on this cut-off score may be less inclined to conclude a diagnosis without knowledge of the decreased stability of the RRB domain, specifically for females, based upon results from this and other studies demonstrating lower correlation values, limited and inconsistent factor loadings, and poorer test-retest values. Clinicians should consider the importance of parent reported RRB symptoms as observed across context and over time, and other supplemental tools that assess these behaviors specifically, and combine that information with observations made during an ADOS-2 administration when evaluating young children with limited verbal language, especially females. Diagnoses of ASD should still be considered even when algorithm cut-off scores are not met within this population when limited RRB symptoms are to blame for that sub-threshold score.

Knowledge of the importance of specific algorithm items within the preschool age developmental period and across sex is another important implication for clinical practice. Past

research has demonstrated the strength at which different items predict ASD diagnosis across language level and developmental period as measured by ADOS-2 Modules. Data presented in this study highlighting the rank order of items for preschool age students and the slight sex differences in ranking are particularly important to consider when working with this age group, which is the group at which most children receive their first developmental evaluation and diagnosis of ASD. When comparing these results to other studies with younger and older children, a consistent picture may also be presented which demonstrates the developmental course of ASD symptoms, information that is critical not only for diagnostic decisions but when identifying targets for intervention.

Limitations

There are many important limitations of this study to consider. First, and perhaps most obvious, the sample was primarily male, perpetuating the significant limitation of most studies exploring the ADOS-2. While other characteristics of the sample were likely more diverse than previous research, including socio-economic status, race, and ethnicity, this sample was 76% male. Although statistical analysis identified the female group as appropriate for factor analysis, the greater *N* of the male group presented with stronger statistical power, thus leaving room for questioning the direct comparison of the results across groups.

In addition, the limitation of this study as a secondary data base analysis implies that other participant characteristics were not available for analysis, for example, cognitive ability and the possible presence of other comorbid diagnoses. These variables may have contributed to the results and would have been important to study across homogenous cells.

Another limitation of this study includes the absence of regular inter-rater reliability assessments in the administration and scoring of the ADOS-2. Though each clinician achieved

reliability through the same rigorous process, checks for reliability were not conducted throughout.

REFERENCES

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York, NY: Macmillan.

APA (1980) *Diagnostic and Statistical Manual of Mental Disorders, 3rd Edition* (DSM-III). Washington, DC: American Psychiatric Association.

APA (1987) *Diagnostic and Statistical Manual of Mental Disorders, 3rd Edition-Revised* (DSM-III-R). Washington, DC: American Psychiatric Association.

APA (1994) *Diagnostic and Statistical Manual of Mental Disorders, 4th Edition* (DSM-IV). Washington, DC: American Psychiatric Association.

APA (2013) *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition* (DSM-5). Washington, DC: American Psychiatric Association.

Bai, D., Yip, B. H. K., Windham, G. C., Sourander, A., Francis, R., Yoffe, R., ... & Gissler, M. (2019). Association of genetic and environmental factors with autism in a 5-country cohort. *JAMA psychiatry*, *76*(10), 1035-1043.

Baron-Cohen, S. (2002). The extreme male brain theory of autism. *Trends in cognitive sciences*, *6*(6), 248-254.

Begeer, S., Mandell, D., Wijnker-Holmes, B., Venderbosch, S., Rem, D., Stekelenburg, F., & Koot, H. M. (2013). Sex differences in the timing of identification among children and adults with autism spectrum disorders. *Journal of autism and developmental disorders*, *43*(5), 1151-1156.

Beggiato, A., Peyre, H., Maruani, A., Scheid, I., Rastam, M., Amsellem, F., ... & Delorme, R. (2017). Gender differences in autism spectrum disorders: Divergence among specific core symptoms. *Autism Research*, *10*(4), 680-689.

Bishop, D. V., & Norbury, C. F. (2002). Exploring the borderlands of autistic disorder and

specific language impairment: a study using standardised diagnostic instruments. *Journal of child Psychology and Psychiatry*, *43*(7), 917-929.

Bishop, S. L., Havdahl, K. A., Huerta, M., & Lord, C. (2016). Subdimensions of social-communication impairment in autism spectrum disorder. *Journal of Child Psychology and Psychiatry*, *57*(8), 909-916.

Boyd, B. A., Baranek, G. T., Sideris, J., Poe, M. D., Watson, L. R., Patten, E., & Miller, H. (2010). Sensory features and repetitive behaviors in children with autism and developmental delays. *Autism Research*, *3*(2), 78-87.

Brown, R. T., Reynolds, C. R., & Whitaker, J. S. (1999). Bias in mental testing since Bias in Mental Testing. *School Psychology Quarterly*, *14*(3), 208.

Carter, A. S., Black, D. O., Tewani, S., Connolly, C. E., Kadlec, M. B., & Tager-Flusberg, H. (2007). Sex differences in toddlers with autism spectrum disorders. *Journal of autism and developmental disorders*, *37*(1), 86-97.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 245-276.

Chawarska, K., Shic, F., Macari, S., Campbell, D. J., Brian, J., Landa, R., ... & Bryson, S. (2014). 18-month predictors of later outcomes in younger siblings of children with autism spectrum disorder: a baby siblings research consortium study. *Journal of the American Academy of Child & Adolescent Psychiatry*, *53*(12), 1317-1327.

Cohen, D., Pichard, N., Tordjman, S., Baumann, C., Burglen, L., Excoffier, E., ... & Héron, D. (2005). Specific genetic disorders and autism: clinical contribution towards their identification. *Journal of autism and developmental disorders*, *35*(1), 103-116.

Constantino, J. N., Todorov, A., Hilton, C., Law, P., Zhang, Y., Molloy, E., Fitzgerald, R., &

Geschwind, D. (2013). Autism recurrence in half siblings: strong support for genetic mechanisms of transmission in ASD. *Molecular psychiatry*, *18*(2), 137–138.

Christensen, J., Grønborg, T. K., Sørensen, M. J., Schendel, D., Parner, E. T., Pedersen, L. H., & Vestergaard, M. (2013). Prenatal valproate exposure and risk of autism spectrum disorders and childhood autism. *Jama*, *309*(16), 1696-1703.

Dean, M., Harwood, R., & Kasari, C. (2017). The art of camouflage: Gender differences in the social behaviors of girls and boys with autism spectrum disorder. *Autism*, *21*(6), 678-689.

De Bildt, A., Sytema, S., Ketelaars, C., Kraijer, D., Mulder, E., Volkmar, F., & Minderaa, R. (2004). Interrelationship between autism diagnostic observation schedule-generic (ADOS-G), autism diagnostic interview-revised (ADI-R), and the diagnostic and statistical manual of mental disorders (DSM-IV-TR) classification in children and adolescents with mental retardation. *Journal of autism and developmental disorders*, *34*(2), 129-137.

DiLavore, P., Lord, C., & Rutter, M. (1995). Pre-Linguistic Autism Diagnostic Observation Schedule (PL-ADOS). *Journal of Autism and Development Disorders, 25,* 355-379.

Duvekot, J., van der Ende, J., Verhulst, F. C., Slappendel, G., van Daalen, E., Maras, A., & Greaves-Lord, K. (2017). Factors influencing the probability of a diagnosis of autism spectrum disorder in girls versus boys. *Autism*, *21*(6), 646-658.

Dworzynski, K., Ronald, A., Bolton, P., & Happé, F. (2012). How different are girls and boys above and below the diagnostic threshold for autism spectrum disorders?. *Journal of the American Academy of Child & Adolescent Psychiatry*, *51*(8), 788-797.

Elsabbagh, M., Divan, G., Koh, Y. J., Kim, Y. S., Kauchali, S., Marcín, C., Montiel-Nava, C.,

Patel, V., Paula, C.S., Wang, C. & Yasamy, M. T. (2012). Global prevalence of autism and other pervasive developmental disorders. *Autism research*, *5*(3), 160-179.

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), 3-31.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods*, *4*(3), 272.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological assessment*, *7*(3), 286.

Frazier, T. W., Georgiades, S., Bishop, S. L., & Hardan, A. Y. (2014). Behavioral and cognitive characteristics of females and males with autism in the Simons Simplex Collection. *Journal of the American Academy of Child & Adolescent Psychiatry*, *53*(3), 329-340.

Goodwin, L. D., & Leech, N. L. (2006). Understanding correlation: Factors that affect the size of r. *The Journal of Experimental Education*, *74*(3), 249-266.

Gotham, K., Pickles, A., & Lord, C. (2009). Standardizing ADOS scores for a measure of severity in autism spectrum disorders. *Journal of autism and developmental disorders*, *39*(5), 693-705.

Gotham, K., Risi, S., Pickles, A., & Lord, C. (2007). The Autism Diagnostic Observation Schedule: revised algorithms for improved diagnostic validity. *Journal of autism and developmental disorders*, *37*(4), 613.

Gotham, K., Risi, S., Dawson, G., Tager-Flusberg, H., Joseph, R., Carter, A., ... & Sigman, M. (2008). A replication of the Autism Diagnostic Observation Schedule (ADOS) revised

algorithms. *Journal of the American Academy of Child & Adolescent Psychiatry*, *47*(6), 642-651.

Grapel, J. N., Cicchetti, D. V., & Volkmar, F. R. (2015). Sensory features as diagnostic criteria for autism: Sensory features in autism. *The Yale journal of biology and medicine*, *88*(1), 69.

Gray, K. M., Tonge, B. J., & Sweeney, D. J. (2008). Using the Autism Diagnostic Interview-Revised and the Autism Diagnostic Observation Schedule with young children with developmental delay: evaluating diagnostic validity. *Journal of autism and developmental disorders*, *38*(4), 657-667.

Hartley, S. L., & Sikora, D. M. (2009). Sex differences in autism spectrum disorder: an examination of developmental functioning, autistic symptoms, and coexisting behavior problems in toddlers. *Journal of autism and developmental disorders*, *39*(12), 1715.

Hendrickson, A. E., & White, P. O. (1964). Promax: A quick method for rotation to oblique simple structure. *British journal of statistical psychology*, *17*(1), 65-70.

Hiller, R. M., Young, R. L., & Weber, N. (2016). Sex differences in pre-diagnosis concerns for children later diagnosed with autism spectrum disorder. *Autism*, *20*(1), 75-84.

Holtmann, M., Bölte, S., & Poustka, F. (2007). Autism spectrum disorders: Sex differences in autistic behaviour domains and coexisting psychopathology. *Developmental Medicine & Child Neurology*, *49*(5), 361-366.

Hull, L., Petrides, K. V., & Mandy, W. (2020). The female autism phenotype and camouflaging: A narrative review. *Review Journal of Autism and Developmental Disorders*, 1-12.

Jacquemont, S., Coe, B. P., Hersch, M., Duyzend, M. H., Krumm, N., Bergmann, S., ... & Eichler, E. E. (2014). A higher mutational burden in females supports a “female

protective model" in neurodevelopmental disorders. *The American Journal of Human Genetics*, *94*(3), 415-425.

Jane, J. S., Oltmanns, T. F., South, S. C., & Turkheimer, E. (2007). Gender bias in diagnostic criteria for personality disorders: an item response theory analysis. *Journal of Abnormal Psychology*, *116*(1), 166.

Jensen, A. R. (1976). Test bias and construct validity. *Phi Delta Kappan*, *58*, 340– 346.

Jensen, A. R. (1980). Bias in mental testing.

Joseph, R. M., Tager-Flusberg, H., & Lord, C. (2002). Cognitive profiles and social-communicative functioning in children with autism spectrum disorder. *Journal of Child Psychology and Psychiatry*, *43*(6), 807-821.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*, 141-151.

Keith, T. Z., & Reynolds, C. R. (1990). Measurement and design issues in child assessment research. In C. R. Reynolds (Eds.), *Handbook of psychological and educational assessment of children*. New York, NY: Guilford Press.

King, M., & Bearman, P. (2009). Diagnostic change and the increased prevalence of autism. *International journal of epidemiology*, *38*(5), 1224-1234.

Kirkovski, M., Enticott, P. G., & Fitzgerald, P. B. (2013). A review of the role of female gender in autism spectrum disorders. *Journal of autism and developmental disorders*, *43*(11), 2584-2603.

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.

Kopp, S., & Gillberg, C. (1992). Girls with social deficits and learning problems: Autism,

atypical Asperger syndrome or a variant of these conditions. *European Child & Adolescent Psychiatry*, *1*(2), 89-99.

Lindsay, K. A., & Widiger, T. A. (1995). Sex and gender bias in self-report personality disorder inventories: item analysis of the MCMI-II, MMPI, and PDQ-R. *Journal of Personality Assessment*, *65*(1), 1-20.

Loomes, R., Hull, L., & Mandy, W. P. L. (2017). What is the male-to-female ratio in autism spectrum disorder? A systematic review and meta-analysis. *Journal of the American Academy of Child & Adolescent Psychiatry*, *56*(6), 466-474.

Lord, C., Risi, S., DiLavore, P., Shulman, C., Thurm, A., & Pickles, A. (2006). Autism from 2 to 9 years of age. *Archives of general psychiatry*, *63*(6), 694-701.

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., ... & Rutter, M. (2000). The Autism Diagnostic Observation Schedule—Generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of autism and developmental disorders*, *30*(3), 205-223.

Lord, C., Rutter, M., & DiLavore, P., & Risi, S. (1999) *Autism Diagnostic Observation Schedule: Manual.* Los Angeles, CA: Western Psychological Services.

Lord, C., Rutter, M., Goode, S., Heemsbergen, J., Jordan, H., Mawhood, L., & Schopler, E., (1989). Autism Diagnostic Observation Schedule: A standardized observation of communicative and social behavior. *Journal of Autism and Developmental Disorders, 19,* 185-212.

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

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

Maenner, M. J., Shaw, K. A., & Baio, J. (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*(4), 1.

Mandy, W., Chilvers, R., Chowdhury, U., Salter, G., Seigal, A., & Skuse, D. (2012). Sex differences in autism spectrum disorder: evidence from a large sample of children and adolescents. *Journal of autism and developmental disorders*, *42*(7), 1304-1313.

Mattila, M. L., Kielinen, M., Linna, S. L., Jussila, K., Ebeling, H., Bloigu, R., ... & Moilanen, I. (2011). Autism spectrum disorders according to DSM-IV-TR and comparison with DSM-5 draft criteria: an epidemiological study. *Journal of the American academy of child & adolescent psychiatry*, *50*(6), 583-592.

McCrimmon, A., & Rostad, K. (2014). Test Review: Lord, C., Luyster, RJ, Gotham, K., & Guthrie, W.(2012)." Autism Diagnostic Observation Schedule, (ADOS-2) Manual (Part II): Toddler Module." Torrance, CA: Western Psychological Services, 2012. Lord, C., Rutter, M., DiLavore, PC, Risi, S., Gotham, K., & Bishop, S." Autism Diagnostic Observation Schedule, " Torrance, CA: Western Psychological Services, 2012. *Journal of Psychoeducational Assessment*, *32*(1), 88-92.

Modabbernia, A., Velthorst, E., & Reichenberg, A. (2017). Environmental risk factors for autism: an evidence-based review of systematic reviews and meta-analyses. *Molecular autism*, *8*(1), 13.

Mussey, J. L., Ginn, N. C., & Klinger, L. G. (2017). Are males and females with autism

spectrum disorder more similar than we thought?. *Autism*, *21*(6), 733-737.

National Center on Birth Defects and Developmental Disabilities, Centers for Disease Control and Prevention. (2020). Autism Spectrum Disorder: Causes and Risk Factors. Retrieved from https://www.cdc.gov/ncbddd/autism/facts.html

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*(6), 1075-1086.

Oosterling, I., Roos, S., de Bildt, A., Rommelse, N., de Jonge, M., Visser, J., ... & Buitelaar, J. (2010). Improved diagnostic validity of the ADOS revised algorithms: A replication study in an independent sample. *Journal of autism and developmental disorders*, *40*(6), 689-703.

Overton, T., Fielding, C., & de Alba, R. G. (2008). Brief report: Exploratory analysis of the ADOS revised algorithm: Specificity and predictive value with Hispanic children referred for autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *38*(6), 1166-1169.

Pinto-Martin, J. A., Levy, S. E., Feldman, J. F., Lorenz, J. M., Paneth, N., & Whitaker, A. H. (2011). Prevalence of autism spectrum disorder in adolescents born weighing <2000 grams. *Pediatrics*, *128*(5), 883–891.

Ratto, A. B., Kenworthy, L., Yerys, B. E., Bascom, J., Wieckowski, A. T., White, S. W., ... & Scarpa, A. (2018). What about the girls? Sex-based differences in autistic traits and adaptive skills. *Journal of autism and developmental disorders*, *48*(5), 1698-1711.

Ramsay, M. C. (1997). Structural equation modeling and test bias. Paper presented at the annual meeting of the Educational Research Exchange, Texas A&M University, College Station.

Ramsay, M. C., & Reynolds, C. R. (2000). Development of a scientific test: A practical guide. *Handbook of psychological assessment*, 21-42

Reinhardt, V. P., Wetherby, A. M., Schatschneider, C., & Lord, C. (2015). Examination of sex differences in a large sample of young children with autism spectrum disorder and typical development. *Journal of autism and developmental disorders*, *45*(3), 697-706.

Reynolds, C. R. (2000). Methods for detecting and evaluating cultural bias in neuropsychological tests. In E. Fletcher-Janzen, T. L. Strickland, & C. R. Reynolds (Eds.), *Handbook of cross-cultural neuropsychology* (249– 285). New York, NY: Kluwer Academic/Plenum Press.

Reynolds, C. R., Lowe, P. A., & Saenz, A. (1999). The problem of bias in psychological assessment. In T. B. Gutkin (Eds.), *The handbook of school psychology* (3rd ed., 549– 595). New York, NY: Wiley.

Reynolds, C. R., & Suzuki, L. A. (2012). Bias in psychological assessment: An empirical review and recommendations. *Handbook of Psychology, Second Edition*, *10*.

Reynolds, C. R., Willson, V. L., & Chatman, S. P. (1984). Item bias on the 1981 revisions of the Peabody Picture Vocabulary Test using a new method of detecting bias. *Journal of Psychoeducational Assessment*, *2*, 219– 221.

Rosenberg, R. E., Law, J. K., Yenokyan, G., McGready, J., Kaufmann, W. E., & Law, P. A. (2009). Characteristics and concordance of autism spectrum disorders among 277 twin pairs. *Archives of pediatrics & adolescent medicine*, *163*(10), 907-914.

Russell, G., Steer, C., & Golding, J. (2011). Social and demographic factors that influence the diagnosis of autistic spectrum disorders. *Social psychiatry and psychiatric epidemiology*, *46*(12), 1283-1293.

Rutherford, M., McKenzie, K., Johnson, T., Catchpole, C., O'Hare, A., McClure, I., ... & Murray, A. (2016). Gender ratio in a clinical population sample, age of diagnosis and duration of assessment in children and adults with autism spectrum disorder. *Autism*, *20*(5), 628-634.

Sanders, S. J., He, X., Willsey, A. J., Ercan-Sencicek, A. G., Samocha, K. E., Cicek, A. E., ... & Goldberg, A. P. (2015). Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron*, *87*(6), 1215-1233.

Schulz, S. E., & Stevenson, R. A. (2019). Sensory hypersensitivity predicts repetitive behaviours in autistic and typically-developing children. *Autism*, *23*(4), 1028-1041.

Shattuck, P. T., Durkin, M., Maenner, M., Newschaffer, C., Mandell, D. S., Wiggins, L., ... & Baio, J. (2009). Timing of identification among children with an autism spectrum disorder: findings from a population-based surveillance study. *Journal of the American Academy of Child & Adolescent Psychiatry*, *48*(5), 474-483.

Snow, A. V., & Lecavalier, L. (2011). Comparing autism, PDD-NOS, and other developmental disabilities on parent-reported behavior problems: little evidence for ASD subtype validity. *Journal of Autism and Developmental Disorders*, *41*(3), 302-310.

Strömland, K., Nordin, V., Miller, M., Akerström, B., & Gillberg, C. (1994). Autism in t halidomide embryopathy: a population study. *Developmental Medicine & Child Neurology*, *36*(4), 351-356.

Szatmari, P. (2000). The classification of autism, Asperger's syndrome, and pervasive developmental disorder. *The Canadian Journal of Psychiatry*, *45*(8), 731-738.

Szatmari, P., Liu, X. Q., Goldberg, J., Zwaigenbaum, L., Paterson, A. D., Woodbury-Smith, M.,

... & Thompson, A. (2012). Sex differences in repetitive stereotyped behaviors in autism: implications for genetic liability. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *159*(1), 5-12.

Sztainberg, Y., & Zoghbi, H. Y. (2016). Lessons learned from studying syndromic autism spectrum disorders. *Nature neuroscience*, *19*(11), 1408.

Van Wijngaarden-Cremers, P. J., van Eeten, E., Groen, W. B., Van Deurzen, P. A., Oosterling, I. J., & Van der Gaag, R. J. (2014). Gender and age differences in the core triad of impairments in autism spectrum disorders: a systematic review and meta-analysis. *Journal of autism and developmental disorders*, *44*(3), 627-635.

Verhoeff, B. (2013). Autism in flux: a history of the concept from Leo Kanner to DSM-5. *History of Psychiatry*, *24*(4), 442-458.

Volkmar, F. R. (1998). Categorical approaches to the diagnosis of autism: An overview of DSM-IV and ICD-10. *Autism*, *2*(1), 45-59.

Volkmar, F. R., Cicchetti, D. V., Cohen, D. J., & Bregman, J. (1992). Brief report: Developmental aspects of DSM-III-R criteria for autism. *Journal of Autism and Developmental Disorders*, *22*(4), 657-662.

Volkmar, F. R., Klin, A., Siegel, B., Szatmari, P., Lord, C., Campbell, M., ... & Buitelaar, J. (1994). Field trial for autistic disorder in DSM-IV. *The American journal of psychiatry*.

Walker, D. R., Thompson, A., Zwaigenbaum, L., Goldberg, J., Bryson, S. E., Mahoney, W. J., ... & Szatmari, P. (2004). Specifying PDD-NOS: a comparison of PDD-NOS, Asperger syndrome, and autism. *Journal of the American Academy of Child & Adolescent Psychiatry*, *43*(2), 172-180.

Widiger, Thomas A. "Invited essay: Sex biases in the diagnosis of personality

disorders." *Journal of Personality Disorders* 12, no. 2 (1998): 95-118.

Wilson, C. E., Murphy, C. M., McAlonan, G., Robertson, D. M., Spain, D., Hayward, H., ... & Zinkstok, J. (2016). Does sex influence the diagnostic evaluation of autism spectrum disorder in adults?. *Autism*, *20*(7), 808-819.

Willson, V. L., Nolan, R. F., Reynolds, C. R., & Kamphaus, R. W. (1989). Race and gender effects on item functioning on the Kaufman Assessment Battery for Children. *Journal of School Psychology, 27*, 289– 296.

Witwer, A. N., & Lecavalier, L. (2008). Examining the validity of autism spectrum disorder subtypes. *Journal of autism and developmental disorders*, *38*(9), 1611-1624.

Zwaigenbaum, L., Bryson, S. E., Szatmari, P., Brian, J., Smith, I. M., Roberts, W., ... & Roncadin, C. (2012). Sex differences in children with autism spectrum disorder identified within a high-risk infant cohort. *Journal of autism and developmental disorders*, *42*(12), 2585-2596.