Title

Autism Spectrum Disorder Variation by Gender: Examining Diagnostic Trends in the Autism Diagnostic Observation Schedule II using Multilevel Modeling and Confirmatory Factor Analysis

Permalink

https://escholarship.org/uc/item/7d76h0nw

Author

Tien, Ingrid Shiyin

Publication Date

2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Autism Spectrum Disorder Variation by Gender: Examining Diagnostic Trends in the Autism

Diagnostic Observation Schedule II using Multilevel Modeling and Confirmatory Factor

Analysis

A thesis submitted in partial satisfaction of requirements for the degree of Masters of Arts in

Education

by

Ingrid Shiyin Tien

2023

ABSTRACT OF THE THESIS

Autism Spectrum Disorder Variation by Gender: Examining Diagnostic Trends in the Autism

Diagnostic Observation Schedule II using Multilevel Modeling and Confirmatory Factor

Analysis

by

Ingrid Shiyin Tien

Masters of Arts in Education

University of California, Los Angeles, 2023

Professor Jeffrey J. Wood, Chair

Approximately 3-4 boys for every girl meet the clinical criteria for autism spectrum disorder

(ASD) in studies of community diagnostic patterns and in studies of autism using samples of

convenience. However, girls with autism have been hypothesized to be underdiagnosed, possibly

because they may present with differing symptom profiles as compared to boys. This secondary

data analysis used the National Database of Autism Research (NDAR) to examine in what ways

gender, symptom profiles, and age are associated with one another in a gold standard assessment

of autism symptoms, the Autism Diagnostic Observation Schedule II (ADOS-II; Lord, 2012).

ADOS-II scores from 6183 children ages 6-14 years from 78 different studies in NDAR

indicated that age and gender were significant predictors of total algorithm, restrictive and

repetitive behavioral, and social communicative difficulties composite severity scores.

Confirmatory factor analysis also determined that the ADOS-II's algorithmic variables' structure differed between the male and female subsamples, such that a partial metric invariance model showed females responding poorly, accuracy wise, to the current algorithmic structure than males on the restrictive and repetitive behavioral total.

The thesis of Ingrid Shiyin Tien is approved.

Catherine Lord Morrison

Michael H. Seltzer

Jeffrey J. Wood, Committee Chair

University of California, Los Angeles

2023

Table of Contents

Autism Spectrum Disorder Variation by Gender: Examining Diagnostic Trends in the Autism Diagnostic Observation Schedule II using Multilevel Modeling and Confirmatory Factor Analysis

Similar to the vast majority of research on neurodevelopmental disorders, autism[1] has been predominantly studied in males[2]. This is unsurprising, considering that males were prominently represented in early autism research conducted by Kanner (1943) and Asperger (1944). Up to the present, a significantly larger ratio of males than females has been reported in both epidemiological samples of community-based autism diagnoses in the US as well as in samples of convenience focusing on children and adults with autism (Barbaresi et al., 2005; Bertrand et al., 2001; Windham et al., 2011; Marino et al., 2020). Researchers have suggested that autism spectrum disorder (ASD) may present differently in girls compared to boys and have encouraged the exploration of sex-differential diagnostic criteria (Halladay et al., 2016). In Lai et al. (2015), a meta-analysis of all publications that related to sex, gender, and autism, studies examining the nosological (how autism is defined) and diagnostic (how autism is identified) process and its variation by gender was found to be the topic receiving the last amount of research attention in comparison to other aspects of autism (e.g., intervention, treatment, genetic or etiological causes, etc.). This study investigates patterns of male-female differences in symptom profiles, and the possible moderating role of age in relation to such differences, within

---

[1] Autism and ASD are used interchangeably to denote anyone who identifies as having autism within the context of the citation or study's diagnostic criteria. The use of both person-first and identity-first language are to acknowledge multiple perspectives from autistic advocates and the neurodiversity movement.

[2] Note that I use the term "sex/gender" to be consistent with Lai et al. (2015), given that the concepts of "sex" and "gender" may be ambiguous and have different meanings for different people. For example, "sex" traditionally refers to one's biological sex at birth, whereas "gender" refers to one's identity as male, female, or non-binary. However, these traditional distinctions may not apply to everyone, so I choose not to make the distinction in order to be more inclusive. Additionally, sex and gender were not differentially distinguished in the sample from the original data collect methods.

data from a gold standard diagnostic measure, the Autism Diagnostic Observation Schedule II (ADOS-II, Lord, 2012), using the National Database on Autism Research (NDAR).

The gender distribution has been observed to be imbalanced, as there are between 3 to 4 boys for every girl who meets the clinical criteria for ASD (cf. Lai et a., 2015; Loomes et al., 2017). It is important to note that most estimates of the male-to-female ratio in autism have been informed by convenience sampling, which is prone to bias of those who receive a community-based diagnosis in the first place. Therefore, the true ratio of males to females with autism in the population is difficult to estimate as studies utilizing epidemiological sampling or sampling the population have had varied results in attempting to establish true rates and because there may be somewhat different symptom presentations among females (Elsabbagh et al., 2012; Fombonne, 2003a). Additionally, the ratio changes with varying levels of cognitive ability - such that with the absence of a comorbid intellectual disability, it has been reported to be 9 males to 1 1 female as compared to 3 or 4 females to 1 female for moderate or severe intellectual disability (Rivet and Matson, 2011). The changing gender ratio with cognitive ability has been identified as a result of biased genetic mutations (Ronbemus et al., 2014), such that females are more likely to present with genetic mutations that are associated with intellectual disability. Supporting this, several studies have also found that autistic females were more likely to have a lower IQ than autistic males (Lord et al., 1982; Lord and Schopler, 1985). Girls with a current autism diagnosis were also found to be more likely than boys with autism to be diagnosed with a past speech problem (Stacy et al., 2014). Thus, the gender ratio is inherently implicated differently for youth without a comorbid intellectual disability.

While there have been several explorations as to how youth with autism differ by gender, reviewed below, little research has been done to explore the nature of gender differences in

autism that might illuminate some of the reasons for this uneven male-to-female diagnostic ratio. It has been argued that this ratio could prove to be a self-fulfilling prophecy (Ratto, 2021), as focusing on males more within the research samples and building the diagnostic profile around males would lead to more males being diagnosed. Therefore, a close examination of gendered expressions of autism symptoms may be informative in better understanding this ratio and, potentially, refining diagnostic practices.

Differing Autism Symptom Profiles in Males and Females

Females have been reported to have a differing symptom profile since earlier explorations (Lord et al., 1982). The autism diagnostic criteria specify measuring restrictive and repetitive behaviors (RRBs) and social communicative difficulties (SCD), of which females have been shown to report lesser RRB symptoms. Amongst children 3-18, autistic girls have a lower reported rate of RRBs but male-level equivalent or even higher levels of social and communicative impairments (Mandy et al., 2012) using parent report and direct observation data. This pattern held true in a study examining the dataset of studies that had administered the Autism Diagnostic Interview (ADI-R) on NDAR for school-aged children (Supekar and Menon, 2015). Teachers and parents also reported males having greater externalizing behaviors than females, offering one possible reason for greater diagnostic rates of males in the community (e.g., Mandy et al., 2012). In this study of both males and females ages 3-18, teachers were also less likely to report difficulties with psychopathology and adaptation for autistic females compared to males who had similar serious social communication difficulties. Parents reports on difficulties were found to have no differences in reciprocal social interaction or communication, but boys were reported to have more RRBs than girls. Additionally, females are found to have lower (McLennan et al., 1993) or similar (Lai et al., 2011) rates of internalizing problems.

In a separate line of research, girls who were found to be less likely to meet the screening cutoff on the Childhood Autism Spectrum Test (CAST) and the subsequent diagnostic criteria on the Development and Wellbeing Assessment (DAWBA), which are used to test mild autism symptoms and psychopathology respectively. This suggests a possible flaw, unique to females, in the diagnostic tools used to detect autism characteristics (Constantino and Charman, 2012). Utilizing the Repetitive Behavior Scale-Revised (RBS-R), the measure's items were found to have a greater success in correctly classifying affected boys (67.9%) than girls (61%; Antezana et al., 2018). Specifically, items that best-discriminated gender were the RRB related items.

Females reporting differing responses in diagnostic assessments have additionally been examined and supported within the context of the ADOS-II as well, as boys have been shown to receive more severe RRB scores than girls on the ADOS-II (Kaat et al., 2020). In this same study, adding sex significantly improved model fit for linear mixed-effects models on half the outcomes in the ADOS-II, indicating that sex affects ASD severity estimates from diagnostic measures among autistic children. Kaat et al. ultimately suggests that there is not a significant enough sex-based severity difference in the performance on gold standard diagnostic tools (including the ADOS-II) to justify a sex-differential diagnostic criteria. On a separate study examining the ADOS-II, females have been found to be less likely to show atypicalities on most items related to social communicative difficulties and on total scores (Rea et al., 2022). However, no sex differences were overall found after adjusting for overall intensity of symptomatology. Even amongst these two articles, there is debate as to whether these findings are clinically significant enough to justify a sex-differential diagnostic criterion – with Rea et al. (2022) suggesting that diagnostic criterias may not have accurately sensitivity to female presentations of autism. Ratto's (2021) commentary directly responding to Kaat et al. (2020)'s

article implores the importance of acknowledging the limitations of relying solely on diagnostic tools to confirm autism diagnoses that relied on fairly homogenous and male-dominated samples, implying a self-fulling prophecy of validation of diagnostic tools. Therefore, Ratto's further recommendation suggested augmenting my current tools with other measures that were developed with inclusion of the full gender spectrum.

      Possible interactions between age and gender. The gender differences in autism symptom profiles seem to widen with age. When examining sex differences in ADOS-II data, it was previously found that the effect size increased from small to moderate as the age shifted from school-aged to adolescence (Kaat et al., 2021). Lower reported rates of RRBs in female versus male autistic youth were found to be the case in children ages 7 to 14 years (Supekar and Menon, 2015) and 3 to 18 years (Mandy et al., 2012) by age-matching the samples by gender; the female phenotype was described as stable across this sample's age range even when controlling for IQ. However, among toddlers with autism, evidence has varied: females 1.5 - 4 years have been found to present significantly less RRB symptoms than males on the ADOS and Child Behavioral Checklist (CBCL; Hartley and Sikora, 2009; Sipes et al., 2011; Szatmari et al., 2012). In contrast, girls were not found to display significantly less RRBs at ages 3 - 4 on the ADOS-II (Harrop, 2015). In adults, females with autism showed more lifetime sensory symptoms and fewer social-communication difficulties (Lai et al., 2011).

Alternative Explanations

      The difference in symptom presentations and prevalence of autism in females might suggest a biological mechanism in which females have a naturally lower prevalence of autism. It has been suggested that girls genetically have protective factors that work in conjunction with male risk factors that protect females from the diagnosis of autism in a multiple-threshold

liability model (Werling, 2016). However, there is no conclusive evidence of sex-specific genetic risk factors at this time (Sanders et al., 2015). It is also unclear whether a genetic protective factor would not only change prevalence, but also presentation of autism in females.

Lived Experiences of Females with Autism

Girls have been found to be diagnosed with autism at significantly later ages, as well as experience greater delays between the time of an initial evaluation to receiving a diagnosis and clinical treatment (Begeer et al., 2013; Shattuck et al., 2009; Siklos and Kerns, 2007). As a result, there may be a proportionately larger sample of later-life diagnoses (22-30 years old) for women than for men (Bargiela et al., 2016). In this study (Bargiela et al., 2016), using qualitative interviews to explore lived experiences, autistic women reported that stereotypes of autism consisting of overt and severe communication problems did not account for how well they camouflaged their symptoms, possibly leading clinicians to suggest that they were "shy" or "good."

Camouflaging, or masking, is reported to be anything from internalizing sensory discomfort or suppressing self-stimulating (stimming) behaviors to scripting or rehearsing conversations, such as preparing banks of "stock phrases" for conversation (Hull et al., 2017). Survey data (Baldwin and Costley, 2015) indicated that many adult women with autism reported needing to develop lists of public responses and to "act normal" on behalf of others. Lai et al. (2017) describes camouflaging as the discrepancy between a person's "external" behavioral presentation in social-interpersonal contexts and the person's "internal" status. By examining their social behaviors, girls as young as elementary-aged have been found to utilize masking behaviors to cope with social challenges and blend into their natural social environments (Dean et al., 2016).While masking behaviors may facilitate adaptation to aspects of daily life in some

ways, the women in Bargiela et al. (2016) felt that masking was detrimental towards the recognition of their autism diagnosis, with the possible consequences of burnout and emotional distress.

Notably, camouflaging has been associated with more depressive symptoms overall (Cage and Troxell-Whitman, 2019; Atherton et al., 2021). Perhaps internalized oppression, or the negative self-perceptions that might be expected to accompany suppression of core aspects of the self can lead to reduced agency and damage one's moral identity (Liebow, 2020). Additionally, the effort required to engage in masking from day to day can be catalysts triggering depression in some individuals with autism, many of whom have a predisposition to emotion regulation challenges (Cage et al., 2018a).

Relatedly, young women with autism have reported that social mimicry and "extreme strategies" (e.g., blaming themselves for being bullied; internalizing that they are "rude" or "lazy") were needed to "appear normal" (Bargiela et al., 2016). Women have reported that using superficial adaptive strategies to cope in social situations has often exposed them to risky social situations (Holliday-Willey, 2014; Baldwin and Costley, 2015). For example, having autism as a woman is linked to an increased risk of being sexually abused (Steward, 2014) – with one sample having nine out of the fourteen women with later-life diagnoses in the sample recounting first-hand experiences of sexual abuse, often posited to be explained by their lack of understanding of social nuances (Bargiela et al., 2016).

The experiences of adults with autism informs the importance of earlier knowledge of a diagnosis and adequate interventions. For example, with the use of social skills interventions, females with autism may learn how to positively cope with complex social situations instead of internalizing difficulties that lead to mental health consequences. These unique experiences of

females with later-life diagnoses also benefit the understanding of how symptom profiles at young ages manifest into differing experiences and expressions of autism.

Background on Gender Theory and Its Potential Applications

Gender socialization, often examined through Gender and Social Role Theory, looks at how individuals are taught to behave socially according to their assigned gender. These assigned gender actions are often defined as role assimilation, explored in the ideas of Role Theory (Lopata, 1994). These theories state that there are many agents of gender socialization, which are outside factors that influence the way a child behaves - something even as simple as parents giving their infants different types of toys or playtime. This theory focuses on the explicit and implicit changes to socialized behaviors that females have for performing their gender roles. Girls have been found to be more popular and socially accepted when prescribing to the ideology of precocity (being an accomplished, mature, and well-spoken person) and domesticity (having a caring and home-centered nature) - completely contrasting with boys' gender orientation that is tied to the ideology of autonomy (Adler et al., 1992; Dulla and Priyadarshini, 2021).

Girls with autism have often appeared as more shy, quiet, and having less externalizing behaviors and/or use camouflaging techniques to hide outwardly apparent symptoms (Hull et al., 2020). This especially holds true when examining RRBs (Lai et al., 2011), which is a core feature of the description of autism on the DSM-5 and is present on current diagnostic measures. Therefore, from a theoretical perspective, the pervasive nature of gender roles and gender socialization may directly impact how females present their symptoms. Considering autism has been observed to have a differing symptom profile based on gender, the direct contrast to the ideologies in which females are socialized to their gender roles may contribute to masking outwardly apparent behaviors.

Literature Gaps

The heterogeneous nature of autism inherently requires diagnostic assessments to detect a wide range of severity and symptom profiles to accurately assess most individuals who have autism. Based on the research reviewed above, it is unclear whether nosological criteria and diagnostic assessments of autism are well-calibrated to the possibly unique symptom profiles of females with autism. Although the extant literature establishes differing ratios of community diagnosis and autism research participation for males and females, there are relatively few explorations at a more specific level of the possible reasons for this differing ratio. Several studies, reviewed above, have uncovered possible patterns of symptom profile differences among males and females with autism, but unsurprisingly, findings have been variable and possibly related to the child's age. If more reliable symptom profile differences, at least for some age groups, were found for males and females across samples, such findings could inform potential future research on optimal nosological and assessment criteria to ascertain autism in females.

In utilizing the NDAR database, it is important to acknowledge two limitations: any dataset with only females with a confirmed diagnosis is potentially a small subsection of autistic females (Livingston et al., 2022); therefore, patterns and trends found in the NDAR dataset may only be the "tip of the iceberg," and only broader population-based studies can likely resolve some of the questions about this missing population.

The Current Study

The current study is designed to examine the relationship between gender, age, and symptom profiles for autistic youth ages 6-14 years. This study is a secondary data analysis that pulls from 68 studies using one common measure (the ADOS-II) of autism symptoms. The following research questions and hypotheses are to be addressed:

I. Are there female-male differences in rates of total, RRB, or SCD symptoms at the scale and individual-item level on the ADOS-II for youth with autism included in the NDAR database?

Hypothesis I: It is anticipated that females will have significantly fewer RRB symptoms (at the item and scale level) compared to males. Relatedly, it is hypothesized that males will score significantly higher on the severity tests for RRBs (both the uncalibrated severity scores via individual modules as well as calibrated severity scores across all modules). This would be consistent with the general hypothesis that males and females have different autism profiles. It is not anticipated that females will have significantly fewer SCD symptoms in total, in line with previous findings.

II. Will the child's age moderate the relationship between gender and symptom profiles (as shown in Figure 1)?

Hypothesis II: It is hypothesized that male-female differences in RRB symptoms (at the item and scale level) will increase with age across the age groups evaluated in this study. This would be consistent with the general hypothesis that symptom profiles differ more greatly as the child is older.

III. Does the psychometric structure of the ADOS-II differ for females and males?

Hypothesis III: It is hypothesized that the optimal scoring algorithm on the ADOS-II, based on a confirmatory factor analysis, may differ for females and males. Specifically, it is expected that the structure of the latent restrictive and repetitive behavior score scales may differ for females and males in a confirmatory factor analysis. On the contrary, it is expected for the model to fit better without allowing variance in the factor loadings for SCD totals.

Age at Time of Study

⬇

Gender / Sex ⟶ Symptom Profile and Severity Differences from Assessment
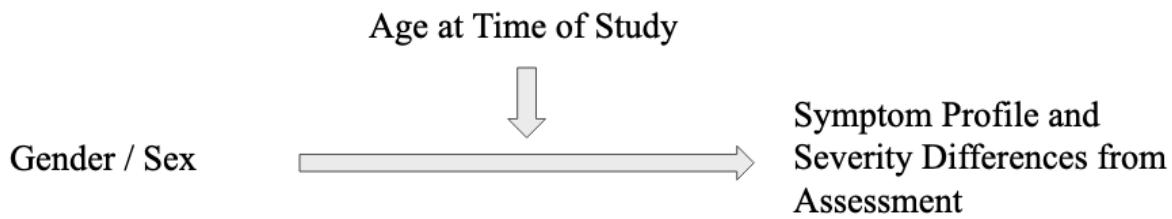
Figure 1. Hypothesized Moderation Effects of Age, Symptom Severity, and Gender

Note. Age at the Time of Study is expected to moderate whether a differing symptom profile would lead to a male-female difference in symptom profiles from a diagnostic assessment.

Methods

Data Source

The participants used in this study were selected from the National Database for Autism Research (NDAR; http://ndar.nih.gov), a National Institutes of Health (NIH) funded data repository that allows access to de-identified data on previously conducted studies on autism spectrum disorders. These participants are drawn from studies with various funding sources using both clinically referred and non-referred participants with an autism spectrum disorder (ASD) diagnosis using the Autism Diagnostic Observation Schedule 2 (ADOS-II; Lord et al., 2012), which served as the primary measure that was examined within the context of this study. These participants came from studies within the United States that had both federal and privately funded grants, as sorted through the inclusion criteria for data being contributed to NDAR. Although different studies used different procedures and assessment instruments to verify autism diagnoses, the inclusion of only studies standardized with the ADOS-II provides a level of quality control with regard to participant recruitment and diagnostic verification.

Data Query

The query output was set to return age, gender, and phenotype along with scores on the ADOS-II. Participants in the current study came from studies that included some form of the diagnostic verification process. The query parameters used with the electronically stored data were between 72 months (6 years) and 168 months (14 years), with a confirmed ASD diagnosis. Only studies with both males and females within the sample were included, with no limitation on sample size in the individual studies selected. The usage of the Autism Diagnostic Observation Schedule II (Lord et al., 2012) must have been present in the study - either as a diagnostic tool, an outcome measure, or both. Data reported for these cases were collected between 2014-2022. Data cleaning processes ensured that participants outside the data range were removed, as well as any cases without complete individual item-level responses. Participants who did not meet the ADOS-II determined criteria for having autism (in comparison to no autism) were also excluded. No other forms of diagnostic verification were included in the dataset – therefore, if participants did not meet diagnostic criteria via the ADOS-II, they were excluded from the study. Next, any participants with any other additional incorrectly entered data (e.g. entering a score of 9 when the subscale only had options 1, 2, or 3) were removed. Lastly, any studies within each sample that did not have at least a sample size of 2, with at least one male and one female, were excluded entirely. See Figure 2 below to examine the full data-cleaning process.

National Database of Autism Research

Using Data Structures Feature to Pull Specific Variables

5687 cases

4100 cases

Obtaining a NIMH Data Archive Data Use Certification

Using the ADOS-II, gender, age

Removing duplicates and missing data

813 females; 3287 males
Aged 6 - 14 years
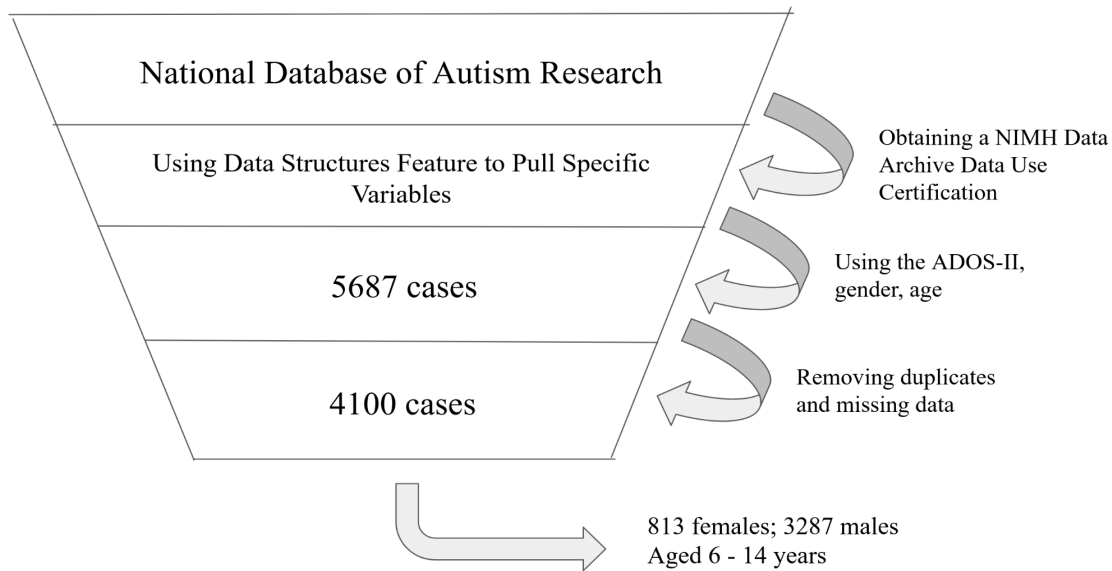
Figure 2. Funnel Figure describing the Data Cleaning and Extraction Process.

Note. Because of both the wide pool of research samples being studied (with the only

commonality being that it is used as a diagnostic measure), as well as the requirement for autism

researchers to submit their data to NDAR, this dataset is representative of all the individuals that

the research community studies. Therefore, this dataset is as inclusive as the research samples

that are included within this secondary data analysis, which acts as a major limitation, as

research samples may not make an inclusive dataset a key goal. However, the primary analysis

of this data also helps draw conclusions on the girls who are currently missed within this

diagnostic process (and therefore do not have any data present), as it is looking at linear trends.

Individual Level Measures

Autism Diagnostic Observation Schedule 2 (ADOS-II; Lord, 2012). The ADOS-II scores

were used to assess symptoms of Autism and their severity, based on the Module that it was

administered in, with higher severity and younger age most likely being administered using

Modules 1 or 2 and lesser severity and/or older age most likely being administered using

Modules 3 or 4. The examiner that is administering the ADOS-II would have chosen which Module to use for each child within the context of each individual study. The administration of the ADOS consists of a series of structured and semi-structured tasks and generally takes around 30 to 60 minutes to administer. During this time, the examiner allows the child to show behaviors as instructed by the manual on behaviors that are relevant to the diagnosis of autism. Subsections of each module all uniformly fall within (A) Language and Communication, (B) Reciprocal Social Interaction, (C) Imagination, (D) Stereotyped Behaviors and Restricted Interests, and (E) Other Abnormal Behaviors - with the weighing of each section and subsection's question being determined by the ADOS-II scoring. Only Modules 3 and 4 have been validated to be reliable at the diagnostic threshold using only the scoring algorithm items, which were the items used within the context of this study (Kuhfeld and Sturm, 2018). This study used both unstandardized (e.g., uncalibrated across modules) severity scores to examine each individual module as well as standardized severity scores to examine the overall dataset. Without calibrated severity scores, algorithm scores were generated to examine an unstandardized severity score to compare amongst the sample within each module. Additionally, the ADOS-II calibrated severity score (ADOS-II CSS) will be used to create a composite severity of autism symptoms for each participant. The ADOS-II was chosen for its diagnostic accuracy and validity, which have a high test-retest reliability and internal consistency (Lord et al., 2000).

Restrictive and Repetitive Behavior Total. The Restrictive and Repetitive Behavior (RRB) Total is an algorithmic score that utilized the scoring manual of the ADOS-II. The items that contribute to this total were Stereotyped/Idiosyncratic Use of Words or Phrases (A4), Unusual Sensory Interest in Play Material/Person (D1), Hand and Finger and Other Complex

14

Mannerisms (D2), and Excessive Interest in Unusual or Highly Specific Topics/Objects or Repetitive Behaviors (D4).

Social Affect Total. The Social Affect (labeled as Social Communicative Difficulties, SCD) Total is the secondary algorithmic score that utilized the scoring manual of the ADOS-II. The items that contribute to this total were Reporting of Events (A7), Conversation (A8), Descriptive, Conventional, Instrumental, or Informational Gestures (A9), Unusual Eye Contact (B1), Facial Expressions Directed to Examiner (B2), Shared Enjoyment in Interaction (B4), Quality of Social Overtures (B7), Quality of Social Response (B9), Amount of Reciprocal Social Communication (B10), and Overall Quality of Rapport (B11).

Age. Age was determined as chronological age, as there was not a standardized understanding of the severity requirements to participate in each study (e.g., I did not have an IQ inclusion criteria), as well as wanting to have a multitude of severity distributions within my final dataset.

Gender. Gender was categorized based on what the original studies reported. The National Database of Autism Research did not allow separate search criteria for sex in contrast to gender. It is important to acknowledge here that not only does one's gender identity not always conform with sex, but in fact, gender-diverse individuals appear more frequently in a population of individuals who have autism or other neurodevelopmental diagnoses (Warrier et al., 2020). However, within the context of this study as a secondary data analysis, due to a lack of understanding of the different study administrations due to NDAR's de-identification system, gender and sex are treated as equivalent constructs in the current study.

Study Level Measures

Study Year. Study year was reported on the NDAR website when pulling each individual study as the last listed end date (when the study was reported to have finished submitting their data).

Sample Size. The sample size was also reported on the NDAR website for each individual study. The sample size listed on the site was reported rather than the sample size within my dataset.

Study Type. Categories were created by examining common themes and trends from abstracts listed on the NDAR website. Study type was coded independently by research assistants by reading the abstracts of the studies. Categories included: (1) genetic study, (2) intervention (both behavioral and medical), (3) neuroimaging, and (4) symptom profiles. Studies could be categorized into multiple types. Interrater reliability amongst coders was acceptance ($\varkappa$ = 0.62).

Study Ratio. Study ratio was reported on the NDAR site and was collected as the percentage of females and males as reported by the original principal investigator on the NDAR site (rather than within my dataset's sample). Study ratio was then utilized as a proportion of females over males. For example, if a study had approximately one female for every 16 males in their study, the study ratio would be 0.0625.

Sample Description

Module 1. Participants in Module 1 were identified from 27 studies, yielding 1577 cases. 420 individuals were parsed for not being in the age parameters, 48 individuals for having an overall total score not indicative of autism, and 25 participants for being in samples with only one participant or only participants of one gender, leaving a final dataset of 1084 participants (319 females, and 791 males). Participants were on average 8.4 years old. Module 1 had both a

"verbal" and "non-verbal" version of the assessment and scoring algorithm; but as this was not information that was privy to the dataset that was received, Module 1 was excluded from the overall severity dataset.

Module 2. Participants in Module 2 were identified from 39 different studies. The same data cleaning process was utilized: from an initial sample of 889, 101 individuals were parsed for not being in the age parameters. 282 individuals were then parsed for not having an overall total score or being in a sample with only one participant or only participants of one gender, leaving a final dataset of 506 individuals (130 females and 376 males), with an average age of 8.5 years old.

Module 3. Participants in Module 3 were identified from 68 different studies, yielding 5687 cases. Using the same data cleaning strategies, participants outside of the data range were removed, leaving 5679 cases. Individual item-level responses from the ADOS-II were missing for 75 children and their cases were removed. Additionally, 1392 cases did not have a reported ADOS-II score that would be indicative of an autism or autism spectrum diagnosis, so they were excluded. 23 additional cases presented missing data that were then deleted. Of the remaining 4100 subjects, 813 were female, and 3287 were male. On average, participants were 9.9 years old.

Module 4 Exclusion. After cleaning out incomplete cases, participants in Module 4 were identified from 5 different studies, yielding only 20 cases. Therefore, Module 4 was excluded from the overall severity dataset as well as from analyses.

Comprehensive Severity Dataset. After compiling the severity scores for participants in Module 2 and 3, the two datasets were merged to create one dataset with uniform severity scores.

Therefore, the final sample yielded 4573 children, with 3639 males and 934 females. The average age of youth in this sample was 9.7 years old.

Results

Multilevel Modeling Analytical Plan

The hierarchical linear model (HLM8 Student Version) program (Bryk et al., 1986; Raudenbush and Congdon, 2021; Hedeker and Gibbons, 2021) was used to fit a series of multilevel models. The HLM program was represented by a hierarchical two-level model by two equations that are estimated simultaneously: the within-study and between-study equations. The within-study model regresses individual ADOS-II total composite scores, social communication difficulty composite scores, and restrictive and repetitive behavioral composite scores on gender and age in months. Gender was added as an uncentered variable and was dummy coded to have males coded as 0 and females as 1. This was in contrast to age, which was added as a group-mean centered variable to provide comparison to the age group within each study. Following this, when significance was found in gender, I used SPSS (Version 28.0) to analyze an interaction effect between gender and age.

Next, in the between-study model, the study ratio (between males and females), study type, sample size, and study year. A series of models were run for the four modules of the ADOS-II, as well as the composite scores represented within modules 2 and 3.

Therefore, the within-study, or child-level model for studies can be written as:

SEVERITY SCORES $= \beta_{0j} + \beta_{1j}(AGE_{ij}) + \beta_{2j}(GENDER_{ij}) + \beta_{3j}(AGE_{ij} * GENDER_{ij}) + r_{ij}$

And the between-study, or individual study level model can be written as:

$\beta_{0j} = \gamma_{00} + u_{0j}$ *(STUDY RATIO)* $+ u_{1j}$ *(SAMPLE SIZE)* $+ u_{2j}$ *(STUDY TYPE)* $+ u_{3j}$ *(STUDY YEAR)*

$\beta_{1j} = \gamma_{10} + u_{1j}$

$\beta_{2j} = \gamma_{20} + u_{2j}$

For each outcome variable (total algorithm, RRB severity, and SCD severity), I ran through the same course of models to examine trends. After reporting the unconditional model, I will be reporting the gender-specific model as well as a model with gender and age controlling for one another. Additionally, to examine further trends by gender, I examined whether the variation in the random slope was significant to infer whether to include a specified random slope for gender within models with study-level characteristics being added. Lastly, I examined the inclusion of each independent secondary study-level characteristic and added in gender in an additional model. Lastly, I attempted to add the combination of primary and secondary level measures that would yield the lowest or least significant random effects variance. Through this, I compared a various number of models with differing random-effects variance to examine which secondary study characteristic made the largest impact in the overall reduction of the random-effects variance, and therefore the amount of variance accounted for within the model.

Multilevel Modeling Results

Total Algorithm Score. Total algorithm scores were created by adding the scores for the items indicated by the ADOS-II that contribute to the overall score for indicating diagnosis for autism. As a result, total algorithm scores are not standardized for severity scores. Therefore, Modules 1, 2, and 3 were compared without the comprehensive module dataset.

Unconditional Model. The unconditional model yielded a significant result for each dataset, indicating the rejection of the null hypothesis. For Module 1, the average total algorithm score was 28.89 for males ($SE = 0.63$) and was significantly different than the null, $p < .001$. The random-effects variance is 7.56 and significant, indicating that the study means of severity score vary significantly across studies, $p < .01$. For Module 2, the average algorithm score for males

was 21.49 (*SE* = 2.56), *p* < .001. The random-effects variance was 102.94 and significant, *p* <

.001. For module 3, the unconditional model resource results showed that the average total

algorithm score for males was 12.44 (*SE* = 31), *p* < .001. The estimation of variance components

was 5.84, *p* < .001.

Gender-only Model. Gender was my first predictor that was examined, which was added to the

model. First, gender proved to be a significant predictor of algorithm scores for Modules 1 and 3,

with marginal significance indicated for Module 2. Moreover, a negative slope for all the

significant slope estimates indicated that males scored higher on average than females. For

Module 1, the average algorithm score for males was 29.17 (*SE* = .65) with a slope estimate of -

1.22 (*SE* = .52), indicating the average score difference between males and females, which was

significant, *p* = .02. The random-effects variance was 7.86 and was significant, *p* < .001. For

Module 2, the average algorithm score was 22.04 (*SE* = 2.57) with a slope estimate of -1.69 (*SE*

= 0.88), *p* = .055. The random effects variance was 102.49, *p* < .001. From module 3, the

average algorithm scores for males was 12.74 (*SE* = .32), *p* < .001. The gender slope estimate

was -1.6, (*SE* = .18), which was also significant, *p* < .001. The estimate of the random effects

variance was 5.84, and was significant, *p* < .001.

Gender and Age Model. Therefore, I began to examine whether there was a significance

difference in age, as well as whether the gender differences would hold while controlling for age.

In Module 1, the average algorithm score for boys who are at the average age of the studies they

belonged to e was 29.20, *SE* = 0.65, *p* < .001. Controlling for gender, the slope estimate for age

was 0.03, *SE* = .01, and was significant, *p* = .001. Lastly, when controlling for age, the gender

slope (-1.35, *SE* = 0.52) was significant, *p* = .010. For Module 2, the average total algorithm

score for boys was 22.10, *SE* = 2.57, *p* < .001. Slope estimate for age was - change in 1 unit in

age in month, .03 change in algorithm score, controlling for gender, $p = .05$. For module 3 the average total algorithm score for boys was 12.74, $SE = 0.32$, $p < .001$. Slope estimate for age was insignificant at -.006, $SE = .003$, $p = 08$. When controlling for age the gender slope estimate was significant at -1.63, $SE = .18$, $p < .001$. The random effects variance still was significant at 5.84, $p < .001$.

Table 1: Total Algorithm Score Gender Differences, when controlling for age

	Female; M (SE)	Male; M (SE)
Module 1*	27.95 (0.52)	29.20 (0.65)
Module 2*	20.22 (0.88)	22.10 (2.57)
Module 3*	11.11 (0.18)	12.74 (0.32)

Denotes significance found

Gender and Age Interaction. Lastly, I examined whether there was an interaction between age and gender for each dataset. In Module 1, there was a significant interaction, $f(1083) = 1.58$, $p = .003$. In Module 2, there was not a significant interaction, $f(505) = 1.05$, $p = .39$. Lastly, in Module 3, there was a significant interaction, $f(4066) = 1.38$, $p = .009$ (see in Figure 3).
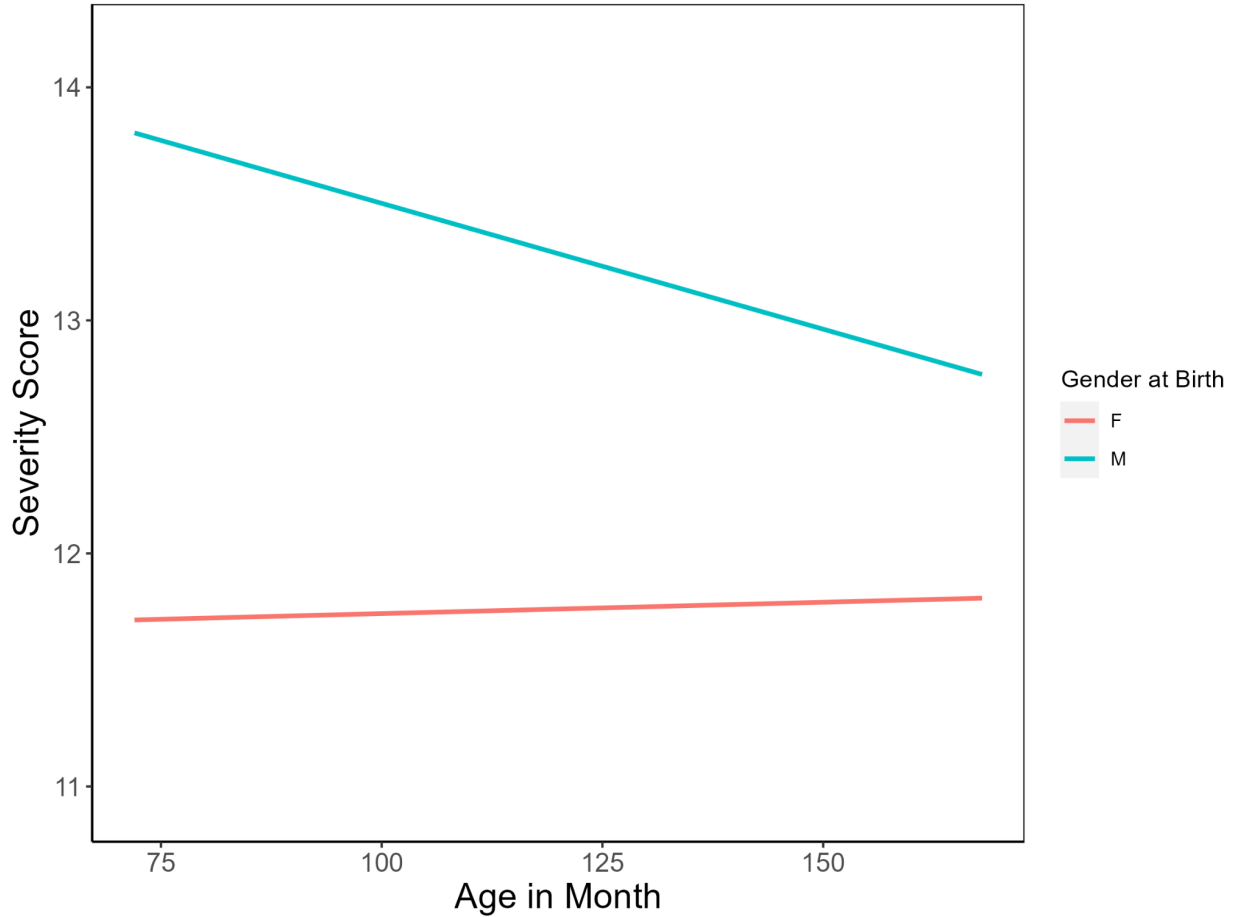
Figure 3. Significant interaction between males and females by age for total algorithm severity score.

Note. Females scored significantly lower than males on all three outcomes variables for most modules, and was found to have a significant interaction by gender with age for multiple outcome variables. Males are seen following a typical trajectory of having a lower severity per age at diagnostic timepoint, which does not follow the trend set by the female sample.

Gender Slope Model. Additionally, I examined whether the variation in the random slope was significant, to infer whether a specific random slope for gender was necessary for study-level characteristics being added. However, there was no significant variation within each module,

leading me to exclude the specific random slope when analyzing the secondary study characteristics.

Secondary Study Characteristics. Therefore, I began examining whether adding any of the additional secondary study characteristics would yield to a significant intercept or a lowering in random effects variance. However, for Modules 1 and 2, there were no significant secondary study characters. Only study type for Module 3 was significant, such that genetic studies and studies looking at symptom profiles are significantly different from one another. The coefficient for the slope estimate was 3.69 ($SE = 1.39$), $p = .01$, even while controlling for age. The random effects variance was 5.14 and remained significant at $p < .001$.

Final Model. Therefore, the final model was estimated based on what had decreased the random effects variance the most. With limited findings of significance of secondary study characteristics, I estimated the final model with age, gender, sample size, and study type. The random effect variance for Module 1 was still significant (7.47), $p < .001$. For module 2, the random effects variance was still significant (126.58), $p < .001$. For module 3, the random effects variance still proved to be significant at 5.81, $p < .001$. Ultimately, in comparison to the unconditional model, there were minimal differences between the random effects variance, indicating little impact as a result of inclusion of the current study variables.

Restricted and Repetitive Behavior.

Unconditional. An unconditional model showed significance with all four datasets. Module 1 presented an average RRB score of 5.50 ($SE = 0.15$), $p < .001$. The random effects variance was significant, .39, $p < .001$. For module 2 the average RRB score was 13.59 ($SE = .11$) $p < .001$. The random effects variance was .26 and was still significant, $p < .001$. Module 3 presented an average RRB score for males of 3.10 ($SE = .1$), $p < .001$. The random effects variance was still

significant (.59), $p < .001$. For the composite data set, the unconditional model showed that on average had a score of 7.44 ($SE =.09$), which was also significant, $p < .001$.

Gender only model. I then continued with the same analytical plan and added gender as a within-study predictor in the model. Gender was found to be a significant predictor in all four datasets. When controlling for gender, the slope estimate for Module 1 (-1.08, $SE = .13$) was found to be significant, $p < .001$. The random-effects variance (0.39) stayed significant, $p < .001$. For module 2, the slope estimate for gender was significant at - 0.78 ($SE = .20$), $p < .001$. The random effects variance was 0.18, and was significant, $p < .001$. The slope estimate for module 3 was -0.31 ($SE = .07$), and was significant, $p <.001$. The random effects variance was also significant at 59, $p < .001$. In the composite data set, gender slope estimate was significant at - 0.25 ($SE = 0.09$), $p = 010$. The final estimation of variance components showed that the random effects variance was still significant at 0.37, $p < .001$.

Gender and Age. When examining a model controlling for both gender and age, the results varied. The slope estimate for age (.005, $SE = .003$) was not found to be significant in Module 1, $p = .08$. However, the slope estimate for gender (-1.10, $SE = .13$) when controlling for age was significant, $p < .001$. The random effects variance (0.39) remained significant, $p < .001$. For module 2, the slope estimate for age was 0.007 ($SE = .004$), $p = 09$. Meanwhile, the gender slope significant at **-0.82** ($SE = 0.20$), $p < 01$. The random effects variance was still significant at 18, $p < .001$. For module 3, the slope estimate for age (-.006, $SE = .001$) was significant, $p < .001$. When controlling for age the gender slope estimate was also significant at -0.31 ($SE = .07$), $p < .001$. The final estimation of random effects variance was also significant at .59, $p < .001$. From the composite data set, the age slope estimate was not significant at -.003, $p = 0.17$. The

gender slope estimate was still significant- 0.25, $SE = 0.09$, $p = 0.01$. The random effects

variance remained significant at 0.37, $p < .001$.

Table 2: Restrictive and Repetitive Behavior Score Gender Differences, when controlling

for age

	Female; M (SD)	Male; M (SD)
Module 1*	4.65 (0.13)	5.75 (0.15)
Module 2*	4.34 (0.20)	5.16 (0.17)
Module 3*	2.53 (0.17)	2.84 (0.07)
Comprehensive Module*	7.24 (0.10)	7.49 (0.09)

Gender + Age Interaction. Therefore, I looked at the interaction between gender and age. In

Module 1, I found that they had a significant interaction, $f(1083) = 3.49$, $p < .001$. In Module 2,

I did not find a significant interaction, $f(505) = 1.05$, $p = .39$. In Module 2, I also found a

significant interaction, $f(505) = 2.44$, $p < .001$. Lastly, in Module 3, I did not find a significant

interaction, $f(4066) = 1.20$, $p = .09$.

Gender Random Slope Model. Similarly, considering the significance of gender within the

model, I examined whether the variation in gender's slope was significant to use in the models

examining secondary characteristics. However, none of the slopes for all four modules varied

significantly, leading to its exclusion in upcoming models.

Secondary Study Characteristics. When estimating secondary study characteristics, it was found

that none of the four study types significantly differed from one another when it came to

restrictive and repetitive behavioral scores for Modules 1 and 3. For module 2, when controlling

for gender, it was found that genetic studies and studies examining symptoms or assessments

were significantly different at - 1.50 difference in score ($SE = 0.54$), $p = .016$. Additionally, it was found that Neuroimaging studies was significantly different than studies examining symptoms or assessments, - 1.21, $SE = 0.46$, $p = 0.02$. The final estimation of random effects variance was .099, which was still significant at $p = .014$. For modules with the composite data set ratio was a significant predictor (1.12), ($SE = 0.36$), $p = 001$. The random effects variance was still significant at 31, $p < .001$. This was still when controlling for gender.

Final Model Estimation. Due to the lack of significant secondary study characteristics, Module 1, and 3 was estimated with age, gender, study type, and sample size. On the contrary, Modules 2 and 4 utilized ratio instead of sample size, as it incurred a decrease in the random effects variance. In Module 1, the random effects variance (0.28) remained significant, $p < .001$. Similarly, in Module 2, the random effects variance was still 0.11, $p = .011$. In Module 3, the random effects variance was also still significant at .58, $p < .001$. In the composite module when running the final model with age, gender, study type, and ratio, the random effects variance was still significant at 28, $p < .001$. Ultimately, the secondary study characteristics did not make a significant impact on the P value of the random effects variance.

Social Communicative Difficulties (Social Affect). The third and final outcome variable was examining the 10 items on the ADOS-II that detected a social affect score.

Unconditional. The unconditional model yielded significance in all four datasets. In module 1, the average algorithm score for boys was 13.59, $SE = 0.11$, $p < .001$. In module 2, the average algorithm score for boys was 12.77, $SE = 0.44$, $p < .001$. In module 3, the average algorithm score for boys was 9.98 ($SE = 0.15$), which was significant, $p < 01$. The composite data set the average algorithm score for boys was 5.85, $p < 005$.

Gender only Model. Gender also proved to be a significant predictor in all four datasets. In module 1 the average algorithm score for boy after controlling for gender was 13.8 ($SE = 0.13$ close from the C, $p < .001$. The slope estimate for gender was also significant at - 0.7 9, $SE = 0.24$, $p = .01$. In module 2, the average algorithm score for boys after controlling for gender was 13.00, $SE = 0.44$). The slope estimate for gender was - 0.80, $SE = 0.41$, which was significant at $p = .050$. The random effects variance was significant at 1.73, $p < .001$. In module 3 the average algorithm score for a boy after controlling for gender was 10.24 ($SE = 0.15$), $p < .001$. The slope estimate for gender was -1.43 ($SE = .15$), $p < .001$. The final estimation of random effects variance was 0.98 and was still significant, $p < .001$. For the composite data set the average algorithm score after controlling for gender was 5.95, $SE = 0.12$, $p < .001$. The slope estimate for gender was significant at -0.46, $SE = 0.07$, $p < .001$. The random effects variance was still significant at .085, $p < .001$.

Gender and Age Model. The results when controlling for age and gender varied between Modules. In module 1, the average algorithm score for boys was 13.83 ($SE = 0.13$), $p < .001$. The slope estimate for age was not significant (.005, $SE = .005$), $p = 0.36$. The slope estimates for gender, however, was significant (-0.80), $SE = 0.24$, $p < .001$.

In module 2, the average algorithm score for boys was 13.04 ($SE = 0.44$). Both age and gender slope estimates were significant, age being .03 ($SE = .01$), $p = .002$. Meanwhile for gender, the slope estimate was -0.94 ($SE = 0.41$), $p = .022$. The random effects variance was still significant at 1.70, $p < .001$. In module 3, controlling for both age and gender, it was found that the average algorithm score for a male was 10.24 ($SE = 0.15$), $p < .001$. The slope estimate for age was not significant at -0.003, $SE = 0.003$, $p = 0.356$. However, the slope estimate for gender was significant at -1.43, $SE = 0.15$, $p < .001$. The random effects variance was still significant at .98,

$p < .001$. In the composite data set, only the gender slope estimate was significant when controlling for age, - 0.46 ($SE = 0.07$), $p < .001$. The age level estimate was not significant at .001, $SE = .001$, $p = 0.412$. The random facts variance was still significant at 0.85, $p < .001$.

Table 3: Social Affect Score Gender Differences, when controlling for age

	Female; M (SD)	Male; M (SD)
Module 1*	13.03 (0.24)	13.83 (0.13)
Module 2*	12.10 (0.41)	13.04 (0.44)
Module 3*	7.38 (0.30)	8.81 (0.15)
Comprehensive Module*	5.49 (0.07)	5.95 (0.12)

Gender + Age Interaction. Therefore, I looked at the interaction between gender and age. In Module 1, I found that they had a significant interaction, $f(1083) = 1.46$, $p = .01$. In Module 2, I also found a significant interaction, $f(505) = 2.24$, $p < .001$. Lastly, in Module 3, I found a significant interaction, $f(4066) = 1.33$, $p = .017$.

Gender Random Slope Model. As conducted before, I examined whether the variation of the gender slope was significant. Our results varied – as Module 3 yielded significance in slopes. In module 1, the variation of the slope gender was not significant at .006, P > .5. In module 2, the variation of the gender slope is not significant at 0.25, $p = 0.289$. Meanwhile, in module 3, the variation of the gender slope was significant at .85, $p = .001$. This meant that I ran the secondary study characteristics against the sloped model.

Secondary Study Characteristics. In mod 1, study type was a significant predictor at 0.40 ($SE = 0.19$), $p = .039$. The random effects variance was still significant, .99, $p < .001$. Ratio was also found to be a significant predictor, as the intercept was at - 2.84, $SE = 1.24$, $p = .031$, even when

controlling four gender. The random effects variance (0.98) was still significant, $p < .001$. In

module 2, genetic studies and studies examining symptoms or assessments were significantly

different from one another, 2.51 ($SE = 0.80$), $p = 0.008$. The random effects variance was still

significant to 0.67, $p = .003$. There were no other significant secondary study characteristics to

report. No additional secondary study characteristics including study year sample size study type

and ratio yielded a significant result for Module 3. For the composite data set, ratio was again

significant at 0.96, $p = .048$. The gender slope estimate was also significant at - 0.46, $SE =$ .10,

$p < .001$. The random effects variance was significant at .81, $p < .001$.

Final model. For module 1, the final model estimated both age, gender, sample size, study type,

and ratio. In the end the random effects variance was still significant at 0.90, $p < .001$.

In module 2 the final model estimated both age, gender, study type, and ratio. The random

effects variance was still significant at 0.79, $p = 0.002$. The final model estimated both age,

gender, sample size, study year, and study type. For module 3, the random effect variance was

still significant at 1.16, $p < .001$. In the composite data set, the final model estimated both age,

gender, study type, and ratio, and found that the random effects was still significant at .80 with p

$< .001$.

Confirmatory Factor Analysis Analytical Plan

Confirmatory factor analysis (CFA) represents a theory-driven approach to test the a

priori factor structure and goodness-of-fit between competing models. Therefore, I conducted a

CFA to examine the structure of the algorithmic severity scores within the ADOS-II and whether

there was any variation by gender. This may suggest differences in the structure of items of

"severity" for boys and girls who already have a diagnosis. This will also inform whether there

are specific domains (RRB vs. SCD) in which the measure is better at capturing diagnostic

values for girls compared with values for boys with autism. This was examined only within

Module 3, as only Module 3 contained enough females in the sample to run a proper analysis.

Distribution Lay-Out

In total, there were 4187 individuals who were included in the dataset analyzing Module

3. The distribution of scores (in Table 4) contributed to the algorithmic Restrictive and

Repetitive Behavior (RRB) and Social Affect (SA) total severity score. As was instructed by the

ADOS-II handbook, scores of 3 were turned into scores of 2, as were scores of 7, 8, and 9 into 0

for the final dataset used to run the CFA.

Table 4: Distribution of Individual Item Responses

Algorithmic Score	Item / Label	0	1	2	3	7	8	9
Restrictive and Repetitive Behaviors	Stereotyped/Idiosyncratic Use of Words or Phrases (A4)	952	2196	868	28	0	0	142
	Unusual Sensory Interest in Play Material/Person (D1)	2237	1132	738	76	0	0	3
	Hand and Finger and Other Complex Mannerisms (D2)	2789	636	624	136	0	0	1
	Excessive Interest in Unusual or Highly Specific Topics/Objects or Repetitive Behaviors (D4)	1105	1087	755	95	0	0	3

Social Affect	Reporting of Events (A7)	1518	1924	605	137	0	0	2
	Conversation (A8)	472	1759	929	45	0	0	0
	Descriptive, Conventional, Instrumental, or Informational Gestures (A9)	1771	1843	334	71	0	0	167
	Unusual Eye Contact (B1)	973	0	3197	0	0	8	7
	Facial Expressions Directed to Examiner (B2)	570	2250	397	0	0	0	8
	Shared Enjoyment in Interaction (B4)	1783	1685	640	78	0	0	0
	Quality of Social Overtures (B7)	320	3014	825	26	0	0	1
	Quality of Social Response (B9)	246	3027	882	29	0	0	2
	Amount of Reciprocal Social Communication (B10)	1288	1987	903	7	0	0	1
	Overall Quality of Rapport (B11)	738	2412	811	71	0	0	154

The configural invariance model, which is the least restrictive, tested whether girls or boys have the same factor structure across groups, with no equality constraints imposed. Therefore, this provided a baseline value for the comparison of a more constrained model. If the data supported the similar factor loading pattern across groups, this model was meant to be tested

against a more restrictive model. I then examined the metric invariance model, which was more restrictive and tested whether each item that contributes to the ADOS-II algorithmic score loaded equivalently onto the same factor for males and females by constraining each item's factor loading to equivalent between each group. If the metric invariance model is not statistically different from the baseline model, then the associations between each item and the overall algorithmic score of the RRB or SCD total would be the same regardless of gender. Once the metric invariance is established, the metric invariance model is tested against even more restrictive model, which constrains that the item means are equal across the gender groups.

Confirmatory Factor Analysis Results

Initially, I examined the overall dataset without differentiating the sample by gender. With this, I found a relatively poor configural invariance model fit with this dataset. The Comparative Fit Index (CFI was .71, the non-normed fit index (TLI) was .65, and the Root Mean Square Error of Approximation (RMSEA) was .01. Therefore, I began to examine ways in which to improve the model fit by differentiating the sample by gender.

Upon examining the configural invariance, I found that the data did not support the same loading pattern across groups, but that this was only the case for certain items. Therefore, upon examining the same dataset using a metric invariance model, it was unsurprising to find that this model was far too restrictive for certain items. When comparing the configural invariance model and the metric invariance model, the chi-square difference was 23.21 with a p-value of .03, indicating a discrepancy between the sample and fitted covariance matrices. The Akaike Information Criterion (AIC) indicated a slight difference between the configural invariance model (115475.5) and metric invariance model (115474.7), as was the case for the Bayesian

Information Criterion (BIC) between the configural invariance model (116020.7) and metric

invariance model (115943.8).

However, considering the improved model fit for the metric invariance model for specific

items, I adjusted the model to freely estimate a subgroup of items across gender groups.

Therefore, in line with my hypothesis, I aligned a partial metric invariance model to examine

whether the items on the RRB totals would benefit from a looser and varied factor loading than

the SCD totals. When examining the goodness of fit between the configural invariance and

partial invariance model, I found varying results indicating a slightly better fit. The chi-square

difference was 18.72 ($p = .07$), indicating a marginally significant discrepancy between the

sample and fitted covariance matrices. Between the AIC, the partial invariance model had a score

of 115472.3, indicating a similar difference between the configural invariance and partial

invariance model. The BIC indicated a similar result, with the partial invariance model having a

value of 115947.7. I also examined the scalar invariance model, where the items means are

constrained to be equal across the groups, and found a significant difference between the metric

invariance and scalar invariance model (87.04, $p < 0.001$). Given the relatively large difference

in model fits, the results from the final model, partial metric invariance model, are reported.

Table 2 displays the standardized factor loadings as well as item means from the final models for

male and female responses for the ADOS-II algorithmic severity score.

Table 5: Factor Loadings and Means by Gender

	Item / Label	Male		Female	
		Factor Loading	M(SE)	Factor Loading	M(SE)
Restrictive and Repetitive Behaviors	Stereotyped/Idiosyncratic Use of Words or Phrases (A4)	1	0.96 (0.01)	1	0.91 (0.02)
	Unusual Sensory Interest in Play Material/Person (D1)	0.93	0.67 (0.01)	0.83	0.60 (.03)
	Hand and Finger and Other Complex Mannerisms (D2)	0.89	0.51 (0.01)	0.53	0.53 (.03)
	Excessive Interest in Unusual or Highly Specific Topics/Objects or Repetitive Behaviors (D4)	0.75	0.69 (0.01)	0.75	0.55 (0.02)
Social Affect	Reporting of Events (A7)	1	0.84 (0.01)	1	0.70 (.03)
	Conversation (A8)	1.18	0.90 (0.01)	1.18	0.83 (.03)
	Descriptive, Conventional, Instrumental, or Informational Gestures (A9)	0.77	0.67 (0.01)	0.77	0.49 (0.02)

Unusual Eye Contact (B1)	0.72	1.56 (0.01)	0.72	1.38 (.03)
Facial Expressions Directed to Examiner (B2)	0.67	0.75 (0.01)	0.67	0.63 (0.02)
Shared Enjoyment in Interaction (B4)	1.30	0.80 (0.01)	1.30	0.53 (0.02)
Quality of Social Overtures (B7)	1.01	1.15 (0.01)	1.01	1.05 (0.02)
Quality of Social Response (B9)	1.03	1.17 (0.01)	1.03	1.10 (0.02)
Amount of Reciprocal Social Communication (B10)	1.41	0.95 (0.01)	1.41	0.76 (.03)
Overall Quality of Rapport (B11)	1.49	1.03 (0.01)	1.49	0.88 (0.02)

Discussion

In line with my hypotheses, boys scored significantly higher than girls on RRB and total algorithm scores, as the slope estimates appeared to be negative. However, it seemed that gender differences were apparent even within SCD scores, indicating gender differences across all sections of the measure. Additionally, only a few secondary study characteristics turned out to have significant variation between studies. These findings suggest that gender differences in ADOS scores are common in many samples. In examining whether the psychometric structure of the ADOS-II severity scores differs between males and females, I found results that supported this. Specifically, I found the largest variation in items that contribute to the RRB behaviors that females have historically presented the lowest rates of. This suggested that the factor structure for the diagnostic algorithm differed somewhat for males and females and that assumptions of scale score equivalence between genders may not be entirely justifiable.

Although this study was not able to examine the initial age of diagnosis, there was some (limited) evidence suggesting boys in some of these samples were younger than girls. However, it is important to consider that some of these research samples used age-matched youth, likely reducing age differences between gender in these studies, and not necessarily reflecting general population trends with regard to age of first diagnosis.

With the use of a secondary dataset, major limitations arise. Ultimately, this study utilized a cross-sectional design that was not able to reflect the referral and recruitment process within each study. Additionally, it is possible that crucial pieces of data were missing from studies' submissions or were inaccurate. Females also comprised a small proportion of the dataset, and therefore, estimates of differences may not be as precise. With this, comes the limitation of a smaller sample size, which was particularly the case in Module 2. Due to this,

multiple findings were particularly skewed and had a large amount of variance within this module for all three outcomes variables. Additionally, a limitation of the CFA was that I was only able to conduct analyses on Module 3, which had enough cases of females to run the analysis.

Overall, these findings provide an understanding of the trajectory or projection of how females tend to appear on the ADOS-II, severity wise, implying that it would be easy for females to be missed in the diagnostic process by not meeting the diagnostic threshold. Specifically, considering that girls demonstrate lesser severity, even when controlling for types of studies, these findings suggest females could be underdiagnosed with the use of structured instruments primarily normed on male samples. However, considering the fluctuation in gender ratio by presence of an intellectual disability, IQ should be further examined within the context of the interactions between these variables. It is important to take into consideration this is likely the tip of the iceberg as to an understanding of how females may be missed in the diagnostic process. Clinician bias and societal stereotypes should be examined to determine how this process occurs, as well as community-based or population-based samples of individuals who are missed, to accurately map the trajectory in which females are underdiagnosed. In conjunction with the CFA results, these findings suggest that there may be a difference in what traits females present that lead to different diagnostic outcomes. Our current diagnostic practices may not be capturing the clinical needs of those who are females and score below the criteria that was determined by mostly male samples when it was derived.

References

Antezana L., Factor R. S., Condy E. E., Strege M. V., Scarpa A., Richey J. A. (2019). Gender differences in restricted and repetitive behaviors and interests in youth with autism. *Autism Research: Official Journal of the International Society for Autism Research*, 12(2), 274–283. https://doi.org/10.1002/aur.2049

Adler, P.A., Kless, S. J., and Adler, P. (1992). Socialization to Gender Roles: Popularity among Elementary School Boys and Girls. *Sociology of Education*, *65*(3), 169–187. https://doi.org/10.2307/2112807

Asperger H.D. (1944). "Autistischen Psychopathen" im Kindesalter. Arch Psychiatry Nervenkr. 117(1), 76–136.

Atherton, G., Edisbury, E., Piovesan, A. et al. Autism Through the Ages: A Mixed Methods Approach to Understanding How Age and Age of Diagnosis Affect Quality of Life. J Autism Dev Disord (2021). https://doi.org/10.1007/s10803-021-05235-x

Baldwin, S., and Costley, D. (2015). The experiences and needs of female adults with high-functioning autism spectrum disorder. Autism: The International Journal of Research and Practice.http://doi.org/10.1177/1362361315590805

Barbaresi, W. J., Katusic, S. K., Colligan, R. C., Weaver, A. L., and Jacobsen, S. J. (2005). The incidence of autism in Olmsted County, Minnesota, 1976-1997: results from a population-based study. *Archives of pediatrics and adolescent medicine*, *159*(1), 37–44. https://doi.org/10.1001/archpedi.159.1.37

Bargiela, S., Steward, R. and Mandy, W. The Experiences of Late-diagnosed Women with Autism

Spectrum Conditions: An Investigation of the Female Autism Phenotype. J Autism Dev Disord 46, 3281–3294 (2016). https://doi.org/10.1007/s10803-016-2872-8

Bertrand, J., Mars, A., Boyle, C., Bove, F., Yeargin-Allsopp, M., and Decoufle, P. (2001). Prevalence of autism in a United States population: the Brick Township, New Jersey, investigation. *Pediatrics*, *108*(5), 1155–1161. https://doi.org/10.1542/peds.108.5.1155

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.

Cage, E., and Troxell-Whitman, Z. (2019). Understanding the Reasons, Contexts and Costs of Camouflaging for Autistic Adults. *Journal of autism and developmental disorders*, *49*(5), 1899–1911. https://doi.org/10.1007/s10803-018-03878-x

Constantino, J., Charman, T. (2012). Gender Bias, Female Resilience, and the Sex Ratio in Autism. *Journal of the American Academy of Child and Adolescent Psychiatry, 51*(8), 756-758. https://doi.org/10.1016/j.jaac.2012.05.017

Dean, M., Harwood, R., and Kasari, C. (2017). The art of camouflage: Gender differences in the social behaviors of girls and boys with autism spectrum disorder. *Autism : the international journal of research and practice*, *21*(6), 678–689. https://doi.org/10.1177/1362361316671845

Dulla, N. and Priyadarshini, S. (2021). Gendered Parenting Influence on Children's Socialization to Gender Stereotype in Marital life. *Rupkatha Journal on Interdisciplinary Studies in Humanities, 13(1)*, 1-19. http://dx.doi.org/10.21659/rupkatha.v13n1.28.

Eaton, N. R., Keyes, K. M., Krueger, R. F., Balsis, S., Skodol, A. E., Markon, K. E., Grant, B. F.,

and Hasin, D. S. (2012). An invariant dimensional liability model of gender differences

in mental disorder prevalence: evidence from a national sample. *Journal of abnormal*

psychology, *121*(1), 282–288. https://doi.org/10.1037/a0024780

Edelson S. M. (2022). Understanding Challenging Behaviors in Autism Spectrum Disorder: A

Multi-Component, Interdisciplinary Model. *Journal of personalized medicine*, *12*(7),

1127. https://doi.org/10.3390/jpm12071127

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., and Fombonne, E. (2012). Global

prevalence of autism and other pervasive developmental disorders. *Autism research :*

official journal of the International Society for Autism Research, *5*(3), 160–179.

https://doi.org/10.1002/aur.239

Fombonne E. (2003). The prevalence of autism. *JAMA*, *289*(1), 87–89.

https://doi.org/10.1001/jama.289.1.87

Geelhand, P., Bernard, P., Klein, O. (2019). The role of gender in the perception of autism

symptom severity and future behavioral development. Molecular Autism 10 (16).

https://doi.org/10.1186/s13229-019-0266-4

Halladay, A.K., Bishop, S., Constantino, J.N., Daniels, A.M., Koenig, K., Palmer, K., Messinger,

D., Pelphrey, K., Sanders, S.J., Singer, A.T., Taylor, J.L., Szatmari, P. (2015). Sex and

gender differences in autism spectrum disorder: summarizing evidence gaps and

identifying emerging areas of priority. Molecular Autism, 6(36). doi: 10.1186/s13229-

015-0019-y

Harrop, C., Gulsrud, A., and Kasari, C. (2015). Does Gender Moderate Core Deficits in ASD?

An Investigation into Restricted and Repetitive Behaviors in Girls and Boys with ASD.

Journal of autism and developmental disorders, 45(11), 3644–3655.

https://doi.org/10.1007/s10803-015-2511-9

Hartley, S.L., and 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, 59(12), 1715–

1722. 10.1007/s10803-009-0810-8.

Hedeker, D. and Gibbons, R, (2021). *SuperMix 3.1: Mixed effects models*. Chapel Hill, NC:

Scientific Software International, Inc.

Holliday-Willey LH. (2014). Pretending to be Normal: Living with Asperger's Syndrome

(Autism Spectrum Disorder) Expanded. Jessica Kingsley Publishers.

Hull L, Petrides KV, Allison C, Smith P, Baron-Cohen S, Lai M-C, et al. (2017b). "Putting on

My Best Normal": social camouflaging in adults with autism spectrum conditions.

Journal of Autism and Developmental Disorders, 47(8), 2519–2534. 10.1007/s10803-

017-3166-5.

Hull, L., Petrides, K.V., Mandy, W. (2020. The Female Autism Phenotype and Camouflaging: a

Narrative Review. *Rev J Autism Dev Disord* **7**, 306–317. https://doi.org/10.1007/s40489-

020-00197-9

IBM Corp. (2022). IBM SPSS Statistics for Windows (Version 28.0) [Computer software]. IBM

Corp.

Kaat, A. J., Shui, A. M., Ghods, S. S., Farmer, C. A., Esler, A. N., Thurm, A., Georgiades, S.,

Kanne, S. M., Lord, C., Kim, Y. S., and Bishop, S. L. (2021). Sex differences in scores

on standardized measures of autism symptoms: a multisite integrative data analysis.

Journal of child psychology and psychiatry, and allied disciplines, *62*(1), 97–106.

https://doi.org/10.1111/jcpp.13242

Kanner L. (1943). Autistic disturbances of affective contact. Nervous Child, 2, 217–50.

Lai, M.C., Lombardo, M.V., Pasco, G., Ruigrok, A.N.V., Wheelwright, S.J. (2011). A

Behavioral Comparison of Male and Female Adults with High Functioning Autism

Spectrum Conditions. PLOS ONE 6(6): e20835.

https://doi.org/10.1371/journal.pone.0020835

Lai, M. C., Lombardo, M. V., Ruigrok, A. N., Chakrabarti, B., Auyeung, B., Szatmari, P.,

Happé, F., Baron-Cohen, S., and MRC AIMS Consortium (2017). Quantifying and

exploring camouflaging in men and women with autism. Autism : the international

journal of research and practice, 21(6), 690–702.

https://doi.org/10.1177/1362361316671012

Liebow, N. (2016). Internalized Oppression and Its Varied Moral Harms: Self-Perceptions of

Reduced Agency and Criminality. *Hypatia, 31*(4), 713-729. doi:10.1111/hypa.12265

Livingston, L.A., Waldren, L.H., Walton, E., Shah, P. (2022). Emotion processing differences

mediate the link between sex and autistic traits in young adulthood, *Journal of Child

Psychology and Psychiatry, 2*(3). https://doi.org/10.1002/jcv2.12096

Lord, C., Schopler, E., and Revicki, D. (1982). Sex differences in autism. *Journal of autism and

developmental disorders*, *12*(4), 317–330. https://doi.org/10.1007/BF01538320

Lord C, Schopler E (1985) Brief report: differences in sex ratios in autism as a function of

measured intelligence. *Journal of Autism and Developmental Disorders* 15: 185–193.

Lord, C., Risi, S., Lambrecht, L. Cook, E., Leventhal, B., DiLavore, P., Pickles, A., Rutter, M.

(2000). The Autism Diagnostic Observation Schedule—Generic: A Standard Measure of

    Social and Communication Deficits Associated with the Spectrum of Autism. J Autism

    Dev Disord 30, 205–223. https://doi.org/10.1023/A:1005592401947

Lord C., Luyster R. J., Gotham K., Guthrie W. (2012). Autism diagnostic observation schedule,

    second edition (ADOS-2) manual (Part II): Toddler module. Torrance, CA: Western

    Psychological Services.

Loomes, R., Hull, L., and Mandy, W. (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 and Adolescent Psychiatry, 56(6), 466–474.

    https://doi.org/10.1016/j.jaac.2017.03.013

Mandy, W., Chilvers, R., Chowdhury, U., Salter, G., Seigal, A., and Skuse, D. (2011). Sex

    differences in autism spectrum disorder: Evidence from a large sample of children and

    adolescents. Journal of Autism Developmental Disorder, 42, 1304–1313.

Marino, F., Chilà, P., Failla, C., Crimi, I., Minutoli, R., Puglisi, A., Arnao, A. A., et al. (2020).

    Tele-Assisted Behavioral Intervention for Families with Children with Autism Spectrum

    Disorders: A Randomized Control Trial. *Brain Sciences*, *10*(9), 649. MDPI AG.

    Retrieved from http://dx.doi.org/10.3390/brainsci10090649

McLennan, J. D., Lord, C., and Schopler, E. (1993). Sex differences in higher functioning people

    with autism. *Journal of Autism and Developmental Disorders, 23*, 217–227.

Ratto, A. B. (2021). Commentary: What's so special about girls on the autism spectrum?

    – a commentary on Kaat et al. (2020). *Journal of Child Psychology and Psychiatry*,

    62(1), 107–109. https://doi.org/10.1111/jcpp.13284

Raudenbush, S.W., and Congdon, R.T. (2021). *HLM 8: Hierarchical linear and nonlinear*

modeling. Chapel Hill, NC: Scientific Software International, Inc.

Rivet, T. T., and Matson, J. L. (2011). Review of gender differences in core symptomatology in

autism spectrum disorders. Research in Autism Spectrum Disorders, 5(3), 957–976.

https://doi.org/10.1016/j.rasd.2010.12.003

Sanders, S.J., He, X., Willsey, A.J., Ercan-Sencicek, A.G., Samocha, K.E., Cicek, A.E., Murtha,

M.T., Bal, V.H., Bishop, S.L., Dong, S., Goldberg, A.P., Jinlu, C., Keaney, J.F., Klei, L.,

Mandell, J.D., Moreno-De-Luca, D., Poultney, C.S., Robinson, E.B., Smith, L., Solli-

Nowlan, T., Su, M.Y., Teran, N.A., Walker, M.F., Werling, D.M., Beaudet, A.L., Cantor,

R.M., Fombonne, E., Geschwind, D.H., Grice, D.E., Lord, C. (2015). Insights into autism

spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron*, 87, 1215–

33. doi: 10.1016/j.neuron.2015.09.016

Shattuck, P. T., Durkin, M., Maenner, M., Newschaffer, C., Mandell, D. S., Wiggins, L., Lee, L.

C., Rice, C., Giarelli, E., Kirby, R., Baio, J., Pinto-Martin, J., and Cuniff, C. (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 and

Adolescent Psychiatry, 48(5), 474–483. https://doi.org/10.1097/CHI.0b013e31819b3848

Sicherman, N., Charite, J., Eyal, G., Janecka, M., Loewenstein, G., Law, K., Lipkin, P.H.,

Marvin, A.R., Buxbaum, J.D. (2021). Clinical signs associated with earlier diagnosis of

children with autism Spectrum disorder. *BMC Pediatrics, 21*(96).

https://doi.org/10.1016/j.ridd.2005.09.003"https://doi.org/10.1186/s12887-021-02551-0

Siklos, S., and Kerns, K. A. (2007). Assessing the diagnostic experiences of a small sample

of parents of children with autism spectrum disorders. *Research in developmental*

disabilities, 28(1), 9–22.

Sipes M, Matson JL, Worley JA, et al. (2011) Gender differences in symptoms of autism

    spectrum disorders in toddlers. *Research in Autism Spectrum Disorders* 5: 1465–1470.

Stacey, M. E., Zablotsky, B., Yarger, H. A., Zimmerman, A., Makia, B., and Lee, L.-C. (2014).

    Sex differences in co-occurring conditions of children with autism spectrum disorders.

    Autism, 18(8), 965–974. https://doi.org/10.1177/1362361313505719

Steward R. (2014). The Independent Woman's Handbook for Super Safe Living on the Autistic

    Spectrum. Jessica Kingsley Publishers

Supekar, K., Menon, V. Sex differences in structural organization of motor systems and their

    dissociable links with repetitive/restricted behaviors in children with autism. Molecular

    Autism 6, 50 (2015). https://doi.org/10.1186/s13229-015-0042-z

Sutherland, R., Hodge, A., Bruck, S., Costley, D., and Klieve, H. (2017). Parent-reported

    differences between school-aged girls and boys on the autism spectrum. Autism : the

    international journal of research and practice, 21(6), 785–794.

    https://doi.org/10.1177/1362361316668653

Szatmari P, Liu X-Q, Goldberg J, et al. (2012) Sex differences in repetitive stereotyped
behaviors

    in autism: implications for genetic liability. *American Journal of Medical Genetics Part
    B: Neuropsychiatric Genetics* 159B, 5–12.

Waizbard-Bartov et al. (2020). Trajectories of Autism Symptom Severity Change during Early

    Childhood. *Journal of Autism and Developmental Disorders*.

Warrier, V., Greenberg, D.M., Weir, E., Buckingham, C., Smith, P., Lai, M.C., Allison, C.,

    Baron-Cohen, S. (2020). Elevated rates of autism, other neurodevelopmental and

    psychiatric diagnoses, and autistic traits in transgender and gender-diverse individuals.

Nat Commun, 11(1):3959. doi: 10.1038/s41467-020-17794-1. PMID: 32770077; PMCID: PMC7415151.

Werling, D.M. (2016). The role of sex-differential biology in risk for autism spectrum disorder. *Biol Sex Differ*, 7(58). doi:10.1186/s13293-016-0112-8

Wilson, C.E., Murphy, C.M., McAlonan, G., Robertson, D.M., Spain, D., Hayward, H., Murphy, D.G. (2016). Does sex influence the diagnostic evaluation of autism spectrum disorder in adults? Autism, 20(7), 808–819. 10.1177/1362361315611381.

Windham, G. C., Anderson, M. C., Croen, L. A., Smith, K. S., Collins, J., and Grether, J. K. (2011). Birth prevalence of autism spectrum disorders in the San Francisco Bay area by demographic and ascertainment source characteristics. *Journal of autism and developmental disorders*, *41*(10), 1362–1372. https://doi.org/10.1007/s10803-010-1160-2