8-2021

Sex Differences in Social Communication Behaviors in Toddlers with Suspected Autism Spectrum Disorder as Assessed by the ADOS-2 Toddler Module

Emily Ronkin
Georgia State University

SEX DIFFERENCES IN SOCIAL COMMUNICATION BEHAVIORS IN TODDLERS WITH SUSPECTED AUTISM SPECTRUM DISORDER AS ASSESSED BY THE ADOS-2 TODDLER MODULE

by

EMILY G. RONKIN

Under the Direction of Erin B. Tone, PhD

ABSTRACT

The Autism Diagnostic Observation Schedule, 2nd-edition (ADOS-2) Toddler Module is the current gold standard measure of autism spectrum disorder (ASD), a neurodevelopmental condition more frequently diagnosed in toddler boys than in toddler girls. Some evidence suggests that the ADOS-2 Toddler Module diagnostic algorithms may capture an ASD phenotype that is more common among toddler boys than toddler girls. Use of these algorithms may thus contribute to observed sex differences in rates of ASD diagnoses. In particular, the diagnostic algorithms give equivalent weight to social communication items on which boys and girls might be expected to score similarly and items on which girls may, as a function of their early socialization histories, perform better than boys. As a consequence, for girls who do have ASD, algorithm scores may inaccurately fail to reach diagnostic cut-offs.

The current study examined the possibility that some ADOS-2 social communication items may function differently for boys and girls by testing the degree to which eight items equivalently related to the social communication latent factor across sexes in a clinical sample (N=315) of toddlers with suspected ASD. Tests of a series of increasingly restrictive models revealed no evidence of sex differences in the current sample, which was inconsistent with hypotheses. Results suggest that the ADOS-2 Toddler Module assesses these eight items in similar ways for boys and girls. Examination of factor loadings point to Creativity/Imagination as a particular area of interest for future research.

INDEX WORDS: Autism spectrum disorder, Toddlers, Sex differences, Social communication, ADOS-2

SEX DIFFERENCES IN SOCIAL COMMUNICATION BEHAVIORS IN TODDLERS WITH SUSPECTED AUTISM SPECTRUM DISORDER AS ASSESSED BY THE ADOS-2 TODDLER MODULE

by

EMILY G. RONKIN

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2020

SEX DIFFERENCES IN SOCIAL COMMUNICATION BEHAVIORS IN TODDLERS WITH SUSPECTED AUTISM SPECTRUM DISORDER AS ASSESSED BY THE ADOS-2 TODDLER MODULE

by

EMILY G. RONKIN

Committee Chair: Erin B. Tone

Committee: Erin C. Tully

Lee Branum-Martin

Lindsey Cohen

Christine Hall

Electronic Version Approved:

Office of Graduate Services

College of Arts and Sciences

Georgia State University

August 2021

ACKNOWLEDGEMENTS

I would like to my friends and family for supporting with me on this journey so far. Finally, thank you to Dr. Erin Tone for reading and editing every page I have sent to you and for your guidance. I eagerly anticipate the future as we continue our work together.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

Autism spectrum disorder (ASD), a neurodevelopmental disorder that affects about 1% of the population, according to recent epidemiological research (Christensen et al., 2016), is more frequently diagnosed in boys than in girls (e.g., Dworzynski, Ronald, Bolton, & Happe, 2012; Whiteley, Todd, Carr, & Shattock, 2010). While ASD is associated with impairments across the lifespan, early intervention has been linked to positive outcomes (e.g., French & Kennedy, 2017); thus, early diagnosis is critical. The current, gold-standard instrument used to assess for ASD in toddlers is the Autism Diagnostic Observation Schedule, 2nd-edition (ADOS-2) Toddler Module (Luyster et al., 2009). During ADOS-2 Toddler Module administration, clinicians observe young children during activities designed to elicit social engagement, play, and verbal or nonverbal communication and document their behaviors. Evaluators enter scores on items from each of five domains (Language and Communication, Reciprocal Social Interaction, Play, Stereotyped Behaviors and Restricted Interests, and Other Behaviors) into diagnostic algorithms that help clinicians and researchers gauge the likelihood that ASD is present.

ADOS-2 Toddler Module diagnostic algorithms, however, may do a better job of assessing male-typical autistic phenotypes than they do phenotypes more common among girls. There are both theoretical and empirical reasons to predict that girls may perform better than boys during ADOS-2 Toddler Module diagnostic observation sessions as a result—at least in part—of gendered socialization of behaviors that items in the social communication domain are designed to capture. Gendered socialization patterns that begin in infancy may modulate the expression and interpretation of ASD symptoms in toddlerhood. Thus, girls with ASD may obtain scores that fail to meet the diagnostic cutoff more frequently than do boys with ASD; this difference may, in turn, contribute to lower ASD rates among girls (e.g., Kreiser & White, 2014).

Most studies to date have examined sex differences in ASD, as measured with the ADOS-2 or similar instruments, at global or domain levels. Such work, however, may obscure patterns that are evident at the level of specific behaviors. Research focused on sex differences at the item level, which could help determine whether selected ADOS-2 Toddler Module items function equivalently or differentially for boys and girls, is only beginning to emerge in the literature (e.g., Beggiato et al., 2017; Wang et al, 2017). Studies that take this approach could provide groundwork for future research into mechanisms that support or produce sex-linked imbalances, as well as for studies that elucidate the presentation of ASD in female toddlers.

The proposed study is designed to assess whether the ADOS-2 Toddler Module diagnostic algorithms comparably assess deficits in boys and girls. To lay a foundation for my hypothesis and research plan, I first provide a brief overview of research on ASD, the ADOS-2 Toddler Module, and sex differences in ASD and its constituent symptoms. I then shift attention to theoretical considerations and empirical evidence that support the idea that specific behaviors develop differently, in both neurotypical toddler boys and girls, as well as those with ASD, at least in part as a result of gendered socialization. This background provides a foundation for my research hypothesis: that specific items in the ADOS-2 Toddler Module (those that measure eye contact, directed affect, social overtures, integration of social behaviors, requesting, shared enjoyment, and imagination/creativity) function differentially for toddler boys and toddler girls, such that boys are more likely than girls to show deficits on these items. Finally, I present the statistical model and plan.

1.1 Autism Spectrum Disorder

ASD symptom expression occurs in two primary domains. The first domain, restricted and repetitive behaviors (RRBs), encompasses intense and exclusive interest in certain topics (e.g., trains, dinosaurs), repetitive motor movements (e.g., hand flapping, spinning, rigid hand and finger posturing) and vocalizations (e.g., repeating words), odd sensory behaviors (e.g., visually inspecting parts of objects, sniffing objects, licking objects), and insistence on sameness/preference for routines (Richler, Bishop, Kleinke, & Lord, 2007). Difficulties in the second domain—social communication—revolve around reciprocally relating to others, recognizing social cues, and communicating nonverbally (Lord, DiLavore, & Gotham, 2012). RRBs and deficits in social communication combine to yield complex functional impairments that are distinctive characteristics of ASD; these include absent or weak pretend play and difficulties communicating with others in social interactions (Charman et al., 1997).

ASD-related anomalies have the potential to disrupt children's social and emotional development and functioning (e.g., Howlin & Magiati, 2017). Further, they often persist into adulthood (Billstedt, Gillberg, & Gillberg, 2007), increasing affected individuals' risk for varied adverse outcomes. Negative sequelae of ASD symptoms include dependence on social services and family support, limited social integration, poor job prospects, and high rates of mental and medical health problems (Bryson & Smith, 1998; Howlin, Goode, Hutton, & Rutter, 2004; Howlin & Magiati, 2017).

There is evidence, however, that early intervention in toddlerhood for neurodevelopmental disorders has long-lasting, positive effects on children's communication, adaptive behavior, and social and emotional functioning (Herskind, Greisen, & Nielsen, 2014; Olson & Montague, 2011). Clinicians have studied and implemented varied treatments for ASD; typically, these have been based on operant learning principles (Lovaas, 1987) and have targeted both toddler communication skills and parent management of child behavior (Boyd, Odom, Humphreys, & Sam, 2010; Matson & Smith, 2008). Behavioral interventions show superiority over other interventions at improving language, behavior, adaptive functioning, pre-academic and social skills, and attention to social stimuli (Dawson et al., 2010; Reichow, Barton, Boyd, & Hume, 2012; Vismara & Lyons, 2007); however, it appears that any early evidence-based intervention is better than no intervention at all, especially in toddlerhood (Boyd et al., 2010; Zwaigenbaum et al., 2015).

Given the long-term implications of ASD and the demonstrated utility of early intervention for improving prognosis, prompt diagnosis is crucial (Zwaigenbaum et al., 2015). Trained clinicians can diagnose ASD in children as young as 12 to 14 months of age (e.g., Mitchell et al., 2006; Pierce et al., 2019). Typically, however, families express initial concerns about ASD symptoms when children are around 3.9 years old (Zablotsky et al., 2017) and ASD is not diagnosed until much later—the average age at diagnosis, despite recent decreases (Fernell & Gillberg, 2010), is currently 5.23 years (Zablotsky et al., 2017). The discrepancy between when clinicians are capable of assigning accurate diagnoses and when children are actually getting diagnosed occurs during the toddler years, a critical period in which early intervention likely has positive effects (Boyd et al., 2010; Zwaigenbaum et al., 2015).

1.1.1 Autism Diagnostic Observation System, 2nd edition (ADOS-2) Toddler Module

The gold-standard assessment tool for ASD in toddlers is the Autism Diagnostic Observation System, 2nd edition Toddler Module (ADOS-2; Falkmer, Anderson, Falkmer, & Horlin, 2013; Luyster et al., 2009), a semi-structured observation scale completed by trained clinicians. The ADOS, which Lord and colleagues (1989) originally developed, has undergone several revisions (DiLavore, Lord, & Rutter, 1995; Gotham, Risi, Pickles, & Lord, 2007; Lord et al., 2000), the most recent of which was in 2012 (ADOS-2; Lord et al., 2012). One of the changes to the ADOS-2 was the inclusion of a Toddler Module (Luyster et al., 2009), which is intended for use in the assessment of children between the ages of 12 and 30 months.

During administration of the ADOS-2 Toddler Module, a trained examiner engages the child in 11 activities over the course of a loosely structured testing session. While the child completes activities that press for social engagement, the examiner observes how the child behaves. Behavioral observations recorded during the assessment provide the basis for determining scores. After administration, examiners score 41 items that tap behaviors likely to occur during module activities. These items fall into five categories (Language and Communication, Reciprocal Social Interaction, Play, Stereotyped Behaviors and Restricted Interests, and Other Behaviors).

Two empirically-derived algorithms combine item scores to yield an Overall Total score on the ADOS-2 Toddler Module (Luyster et al., 2009). The algorithms consist of either 14 or 15 items, depending on the verbal language level of the child (few to no words or some words). These items are pulled from several categories (Language and Communication, Reciprocal Social Interaction, and Stereotyped Behaviors and Restricted Interests), and each is weighted equally. Overall Total scores fall into one of three classification ranges: no concerns of autism, mild concerns, and moderate-to-severe concerns. While the ADOS-2 Toddler Module Overall Total score is an important part of an ASD evaluation, the score is integrated with other information, including clinician judgment, in order to make diagnoses.

Only two published studies, to our knowledge, have focused on the psychometric properties of the ADOS-2 Toddler Module (Guthrie, Swineford, Wetherby, & Lord, 2013; Luyster et al., 2009) and only one conducted a categorical factor analysis of the items (Guthrie et al, 2013). Guthrie and colleagues (2013), using data from a sample of toddlers diagnosed with

ASD, found that the best-fitting factor structure was a two-factor model (Social Communication and Social Interaction [SCI] and Restrictive/Repetitive Language and Behavior [RRLB]), which maps onto the Diagnostic and Statistical Manual, 5th edition (DSM-5; American Psychological Association, 2013) diagnostic criteria.

A similar two-factor structure also emerges for items on the Autism Diagnostic Interview-Revised (ADI-R; Lord, Rutter, & Le Couteur, 1994), a semi-structured interview that trained clinicians administer to parents. Items on the ADI-R, based on parent report of behaviors of toddlers with suspected ASD, cluster into social communication and RRB factors (Kim & Lord, 2012). Notably, two more recent studies examining items from assessments of toddlers with suspected ASD found that a three-factor model also fit the data (Beuker et al., 2013; de Bildt et al., 2015). Beuker and colleagues (2013) found that items on several parent-report measures of autistic symptoms fit into three factors (Social Interaction, Communication, and Stereotyped and Rigid Patterns of Behavior); de Bildt and colleagues (2015) found three slightly different factors (Social Affect; RRBs; and Imitation, Gestures, and Play) when examining items on the ADI-R. Unfortunately, neither study compared fit between the three- and two-factor models, so it is unknown if the inclusion of a third factor significantly improved fit (Beuker et al., 2013; de Bildt et al., 2015).

There are some limitations to Guthrie and colleagues' (2013) published factor analysis of the ADOS-2 Toddler Module. The first is that the researchers did not compare ADOS-2 psychometric properties between male and female toddlers. Second, rather than recruiting a community sample of toddlers, Guthrie and colleagues (2013) focused exclusively on toddlers diagnosed with ASD. This decision is not surprising; researchers do not often administer the ADOS-2 Toddler Module to community samples due to the resources (e.g., time, trained clinicians) required for administration. However, as a consequence of this decision, this sample excludes children who do not exhibit hallmark diagnostic indicators of autism, which likely led to an underrepresentation of girls.

The lack of published studies comparing the factor structure of the ADOS-2 Toddler Module between boys and girls leaves a notable gap in the literature and underscores a need for further research examining the measure's psychometric properties. Further, there are marked sex differences in ASD diagnostic rates (e.g., Dworzynski et al., 2012; Whiteley et al., 2010) that are not well understood; research that clarifies whether the ADOS's psychometric properties are variant across the sexes is also warranted.

1.1.2 Sex differences in ASD

Across the lifespan, regardless of the instruments used to assess symptoms, males more frequently receive ASD diagnoses than do females. Ratios range from 2:1 (e.g., Dworzynski et al., 2012) to 12:1 (e.g., Whiteley et al., 2010), and they appear to vary according to intellectual abilities. Whereas studies that sample participants with comorbid intellectual disability (ID) find male to female ratios as low as 2:1 or 3:1 (e.g., Dworzynski et al., 2012; Loomes, Hull, & Mandy, 2017; Shaw et al., 2020), research focused on individuals without comorbid ID has yielded evidence of larger sex differences (7:1; 12:1; Simonoff et al., 2008; Whiteley et al., 2010). Sex differences in ASD diagnosis rates appear to be a remarkably consistent finding (see Giarelli et al., 2010; Kirkovski, Enticott, & Fitzgerald, 2013; Loomes et al., 2017; Rivet & Matson, 2011 for review articles). Fombonne (2003) wrote that he was not aware of a research sample in which girls were diagnosed with ASD more frequently than were boys, and no such findings appear to have emerged in the published literature in the subsequent 17 years.

We know little about the mechanisms that underlie sex differences in the presentation of ASD. Researchers generally agree that ASD is a heritable disorder that may reflect sex-linked patterns of genetic anomalies (e.g., Hicks et al., 2018; Jaquemont et al., 2014; Tsai, Stewart, & August, 1981; Wing, 1981). Epidemiological, epigenetic, and brain imaging findings lend support to genetic hypotheses regarding the etiology of ASD (e.g., Cauvet et al., 2018; Mattila et al., 2011; Skuse, 2000; 2007; Wierenga et al., 2017), as do a few large family heritability studies (e.g., Turner et al., 2019; Werling & Geschwind, 2013). However, genetic theories inadequately account for the sex differences in diagnostic rates and phenotype of ASD as measured by the ADOS-2 Toddler Module. Environmental factors, such as gendered patterns of socialization, may also modulate the expression and interpretation of ASD symptoms in boys and girls.

In line with this possibility, researchers have suggested that the ways in which we differentially socialize girls and boys may play driving roles in the emergence of sex differences in ASD (Hartung & Widiger, 1998; Kreiser & White, 2014). In particular, they raise the possibility that socialization for girls prioritizes behaviors that may help them mask or camouflage ASD-related deficits (Kreiser & White, 2014). Specifically, gendered socialization may lead girls to learn “surface-level” social behaviors (e.g., brief eye contact, a wider range of facial expressions) that enable girls to conceal social communication deficits traditionally characteristic of ASD. Socialization for boys, in contrast, focuses on behaviors that are less likely to serve a camouflaging function. Such skewed socialization processes for children of both genders may contribute, in turn, to biased patterns of diagnosis when clinicians rely on standard ASD assessment tools (Hartung & Widiger, 1998). Specifically, the ADOS-2 Toddler Module measures several behaviors that gendered socialization processes plausibly influences, raising the possibility that it is biased to detect a male-typical ASD phenotype.

Longstanding evidence demonstrates that adults encourage behaviors in girls and boys—via observational modeling, reinforcement/punishment, and direct teaching—that are consistent with socially sanctioned gender roles (Bussey & Bandura, 1999). They thus set the expectation that girls will develop a set of interpersonal, social, and emotional skills that differs from that of boys. Such sex-linked socialization begins in, and even before, infancy. For example, caregivers differentially structure environments for boys and girls; they commonly choose names, clothing, and room decorations based on the observed sex of their child and they also emphasize and value different physical attributes (e.g., strength vs. beauty) in boys and girls (Johnson, Lurye, & Tassinary, 2010; Thompson & Bentler, 1971). A variety of sex-linked socialization processes appear particularly influential for the development of social behaviors in young children that may serve to camouflage ASD-related deficits in girls. These processes cluster roughly into two categories: caregiver-child interactions and manifestation of these interactions in the context of play.

1.1.2.1 Caregiver-child interaction

Caregivers interact differently with boys and girls from infancy through toddlerhood (e.g., Malatesta & Haviland, 1982). Several studies have found that parents express more emotions (Malatesta & Haviland, 1982) and more frequently speak about emotions (Fivush, Brotman, Buckman, & Goodman, 2000) with infant and toddler daughters than with sons. Caregivers are also more likely to reinforce displays of sadness and worry in toddler and preschool-aged girls than in boys (Adams, Kuebli, Boyle, & Fivush, 1995; Garside & Klimes-Dougan, 2002), and to reinforce toddler girls for engaging with others in intimate, social interactions (Goodwin, 2006). In marked contrast, caregivers withhold attention from toddler-aged girls when they express anger or frustration (Garside & Klimes-Dougan, 2002). Thus, it is not surprising that, by 36 months, typically-developing girls both make more references to feelings and express more emotions, such as sadness, worry, and sympathy, than do boys (Brown, Craig, Halberstadt, 2015; Chaplin & Aldao, 2013; Fivush, 1989).

A different pattern emerges during parent interactions with boys. Not only do caregivers withhold attention in reaction to boys' displays of sadness, but they also reinforce toddler and preschool-aged boys for displaying anger and frustration (Adams et al., 1995; Garside & Klimes-Dougan, 2002) and engaging in "rough-and-tumble" active play and aggression with others (DiPietro, 1981; Martin & Ross, 2005). Typically-developing boys tend to behave in accordance with this socialization; by preschool, there is some evidence that they are more likely than girls to express anger (Chaplin & Aldao, 2013), particularly in the presence of strangers (Cole, Zahn-Waxler, & Smith, 1994). Moreover, toddler boys, on average, appear to engage more often than do girls in active play that involves forceful or rough physical contact (Fabes, Martin, & Hanish, 2003).

Caregivers not only differentially reinforce gendered patterns of behavior in children, but they also engage in other actions, particularly in the context of interpersonal dyads, that may lead girls and boys to show distinctive social development trajectories. In several studies, for instance, mothers matched, or mirrored, infant boys' facial expressions more often than they did those of infant girls (Malatesta & Haviland, 1982; Tronick & Cohn, 1989; Weinberg, Tronick, Cohn, & Olson, 1999). Some findings also suggest that caregivers engage in more synchrony, or tightly-linked, reciprocal verbal and nonverbal interactions (Delaherche et al., 2012), with infant boys than with girls (e.g., Tronick & Cohn, 1989; Weinberg et al., 1999). Researchers have posited that by providing girls with more limited mirroring and synchrony from infancy, caregivers require girls to engage more independently with them and to "hold their own" in interactions. The increased mirroring and synchrony that caregivers provide to infant boys, in contrast, scaffolds interactions and prevents boys from taking the lead in structuring social connections (Weinberg et al., 1999). Thus, girls quickly may become more skilled at social interactions, especially those that involve emotions, because they are expected, from an early age, to be more interpersonally self-reliant than are boys.

Parental socialization also appears to influence infants' and toddlers' reactions to social stimuli, such as faces, voices, and gestures (Connellan, Baron-Cohen, Wheelwright, Batki, & Ahluwalia, 2000; Hittelman & Dickes, 1979; Mundy et al., 2007; Osofsky & O'Connell, 1977). For instance, during infancy and toddlerhood girls orient to faces and voices more frequently than do boys (Connellan et al., 2000; Gunnar & Donahue, 1980; Kleberg, Nyström, Bölte, & Falck-Ytter 2018). They also hold eye contact longer (Connellan et al., 2000; Hittelman & Dickes, 1979; Mundy et al., 2007; Osofsky & O'Connell, 1977) and are more likely to attend and respond to social gestures, such as pointing (Harrop et al., 2015; Øien et al., 2016). Not surprisingly then, preschool girls appear, on average, to hold slight advantages over boys in attending to and accurately processing social cues such as facial and vocal expressions and engaging with others in emotionally sensitive ways (Denham, Mitchell-Copeland, Strandberg, Auerbach, & Blair, 1997; McClure, 2000).

There is some evidence that girls on the autism spectrum similarly outperform boys with ASD on emotion production and processing tasks. With regard to emotion production, toddler girls with ASD produce more directed facial expressions than do their male counterparts (Sipes, Matson, Worley, & Kozlowski, 2011), a pattern that is also apparent in older girls with ASD (Beggiato et al., 2017; McLennan, Lord, & Schopler, 1993). Older girls with ASD may also use emotion content in more contexts than do boys. For instance, in two small studies, school-aged girls generated fictional narratives and autobiographical memories that were more emotional and detailed than those that boys with ASD generated (Conlon et al., 2019; Goddard, Dritschel, & Howlin, 2014).

Similar findings are evident with regard to emotion processing; however, all published findings come from studies of adolescents and adults with autism. Few to no studies examine emotion processing in toddlers with ASD, likely due to challenges assessing this construct in preverbal children. Several studies have shown that older girls with ASD are better at attending to and understanding others' facial expressions than are their male counterparts (e.g., Forgeot d'Arc et al., 2017; Hall, Hutton, & Morgan, 2010; Harrop et al., 2018). Additionally, some studies have found autistic traits to correlate significantly and negatively with facial expression recognition in older males, but not females (Kothari, Skuse, Wakefield, & Micali, 2013; Matsuyoshi et al., 2014; Whyte & Scherf, 2018), suggesting that girls' ability to recognize facial expressions may not be as tightly linked to ASD symptomology as it is for boys.

Caregiver behaviors likely influence child behavior; however, the converse is also probable (Pettit & Arsiwalla, 2008). Thus, children and their caregivers interact in dynamic and reciprocal ways that may contribute to the development of gendered behaviors (Bussey & Bandura, 1999). Further, living in contexts where gendered socialization is normative likely influences how caregivers and other adults who interact with children, such as healthcare providers, interpret child behavior. For instance, caregivers and providers who are aware that rates of ASD are elevated among boys might interpret the same behaviors differently in boys than in girls, identifying them as problematic in boys, but not girls (Kreiser & White, 2014; Rivet & Matson, 2011). Such biased interpretations of particular behaviors may contribute to discrepant ASD diagnostic rates (Dworzynski et al., 2012; Whiteley et al., 2010). If adults are indeed biased to expect ASD to occur more often in boys, it is not surprising that parents report more developmental concerns about boys than about girls (Ramsey et al., 2018). Further, clinicians typically notice developmental delays earlier in boys than in girls (Wang et al., 2017) and assign ASD diagnoses to boys at slightly younger ages than girls (Petrou, Parr, & McConachie, 2018; Shattuck et al., 2009).

1.1.2.2 The context of play

One context in which the ways in which caregivers interpret emotional cues and socialize emotional behaviors may vary by sex is in the context of play. From early in development, for instance, caregivers provide different toys and play with young children in subtly different ways, depending on their observed sex (Alexander, Wilcox, & Woods, 2009; Pomerleau, Bolduc, Malcuit, & Cosette, 1990; Rheingold & Cook, 1975). Whereas toddler boys typically receive machines, vehicles, and sports equipment, toddler girls receive baby dolls, doll houses, and domestic items, all of which encourage pretend play focused on social themes and assumption of others' perspectives (Harrop, Green, Hudry, & the PACT Consortium, 2017). Recently, parents have reported increasing efforts to provide children, regardless of their observed sex, with a range of both stereotypically "male" and "female" toys (Freeman, 2007). Children, however, still demonstrate preferences for traditionally gendered toys and activities from as early as 12 months of age, and these preferences remain consistent throughout toddlerhood (Freeman, 2007; Jadva, Hines, & Golombok, 2010; Servin, Bohlin, & Berlin, 1999; Zosuls et al., 2009). Similar findings in non-human primates suggest that such preferences may be evolutionarily grounded (Hassett, Siebert, & Wallen, 2008).

Sex differences also appear in the types of play in which children engage and the ways in which parents interact with children while playing. These sex differences may emerge at least in part because stereotypical "girl toys" (e.g., dolls, toy kitchens) tend to be more conducive to complex and independent pretend play than are "boy toys" (e.g., trucks, balls; Cherney, Kelly-Vance, Gill Glover, Ruane, & Oliver Ryalls, 2003). In addition, caregivers have been shown to provide more explicit guidance during pretend play for toddler and preschool-aged sons than for daughters (Farver & Wimbarti, 1995; Lindsey & Mize, 2001). Specifically, in at least one study, mothers gave more play leads and direct commands to their preschool-aged boys than girls (Lindsey & Mize, 2001). Lindsey and Mize (2001) posited that sex differences in parent-child interactions during play contribute to the development of gender roles, in that parents model for children how boys and girls are expected to behave, demonstrating assertive behavior with boys and cooperative behavior with girls.

Moreover, given that parents tend to provide more explicit guidance, especially during pretend play, for boys than for girls, they may be providing implicit feedback that preschool-aged girls are inherently better at pretend play (Farver & Wimbarti, 1995; Lindsey & Mize, 2001). Not surprisingly, then, typically-developing preschool girls may exhibit more frequent and more complex pretend play than do boys (Lindsey & Mize, 2001; but see Carlson & Taylor, 2005 for contradictory findings).

Gender socialization processes also occur in the context of play for children with ASD (e.g., Harrop et al., 2017), and gendered patterns of play are evident in this population as well. Both parents and independent observers report that toddler girls with ASD—like their typically-developing peers—are more likely to engage in pretend play than are their male counterparts (Campbell et al., 2018). A similar pattern has been observed in older girls with ASD (Beggiato et al., 2017; Hiller, Young, & Weber, 2014; Knickmeyer, Wheelwright, & Baron-Cohen, 2008; McLennan et al., 1993; but also see Stanley & Konstantareas, 2007, who failed to detect sex differences). Additionally, clinicians in one observational study rated toddler and preschool-aged girls with ASD as more engaged with both parents and examiners during pretend play paradigms than were boys (Campbell et al., 2018).

1.1.2.3 Summary

Gendered socialization processes influence the ways in which children behave in both interpersonal and play contexts, as well as the ways in which caregivers and providers perceive and interpret child behavior. Thus, gendered socialization likely interacts with genetic mechanisms to contribute to sex differences in toddler boys' and girls' patterns of social and emotional functioning. For neurotypical individuals, the real-world impact of these socialized sex differences in emotional and social skills appears to be modest (e.g., Li & Wong, 2016). However, for toddlers with ASD, gendered socialization may confer substantial advantages or disadvantages in interpersonal contexts, especially in short diagnostic testing sessions that set up scenarios that pull for social communication behaviors for clinicians to observe.

In particular, toddler girls may, as a function of traditional feminized socialization, acquire interpersonal, social, and emotional skills that help them to camouflage their ASD symptoms more effectively than can toddler boys (Dean, Harwood, & Kasari, 2017). Evidence-based gender socialization models (Hartung & Widiger, 1998; Kreiser & White, 2014) suggest that parents more heavily socialize girls (regardless of ASD status) to engage in behaviors that, at least superficially, signal interpersonal connection (e.g., more emotion production and better emotion processing) and assumption of others' perspectives (e.g., pretend play). Girls with ASD who engage in these social behaviors—even if only in a superficial or stereotyped way—thus appear less socially impaired than do boys with the same condition, who are less likely to show signs of social engagement.

Simultaneously, caregiver and provider expectancy biases influence the ways in which they interpret ASD symptoms. Thus, caregivers and clinicians may interpret the same behaviors differently in boys and girls, which may bias ADOS-2 Toddler Module scores at the item level. As a consequence, clinicians may be less likely to identify and diagnose ASD in girls than in their male peers in toddlerhood, when intervention appears to have particularly valuable effects (Boyd et al., 2010; Zwaigenbaum et al., 2015). To decrease this possibility, clinicians need measures that effectively capture the ASD phenotype for all children flagged for an ASD evaluation. Specifically, a measure that assesses the phenotype equally well in boys and girls in toddlerhood is necessary, given that early diagnosis is associated with later positive effects for communication, behavior, and adaptive skills (Dawson et al., 2010; Reichow et al., 2012; Vismara & Lyons, 2007).

1.2 The Proposed Study and Research Aims

An initial step toward elucidating the role that gendered socialization may play in driving sex differences in ASD as assessed by the ADOS-2 Toddler Module is to examine sex differences in behaviors that may be influenced, at least in part, by gendered socialization processes. Most research on sex differences in ASD to date, however, has focused on omnibus differences in ADOS-2 scores that aggregate gendered and non-gendered behaviors. This approach, which has yielded variable findings, may obscure more finely-grained distinctions between boys and girls. Moreover, many of the studies that examine behavior-level sex differences in the extant literature rely on small samples and dated measures of ASD (e.g., Beggiato et al., 2017; Harrison, Long, Tommet, & Jones, 2017; Hiller et al., 2014; Kopp & Gillberg, 2011; Øien et al., 2016). Thus, studies are needed that examine discrete social behaviors, particularly those that parents differentially reinforce and those that providers may differentially interpret in girls and in boys. Such work would further benefit from the use of data collected using the gold-standard ASD measure for toddlers, the ADOS-2 Toddler Module, in large samples of youths.

In the current study, I aimed to address this gap in the literature by examining sex differences at the item level for behaviors assessed with a gold-standard ASD diagnostic measure, the ADOS-2 Toddler Module (Lord et al., 2012; Luyster et al., 2009), in a sample of toddlers referred by pediatricians for ASD evaluations. More precisely, I used a structural equation modeling (SEM) approach to differential item functioning (DIF) to determine whether selected social communication items in the ADOS-2 Toddler Module related to the latent social communication phenotype with comparable effectiveness in girls and boys.

I chose eight items from the ADOS-2 Toddler Module to include in my model. I selected these eight items based on both theoretical considerations and empirical evidence. I chose to focus on items that tapped into social communication behaviors because, although sex differences in RRBs are robust in children over six years of age, toddler samples show similar rates of RRBs in boys and girls (see van Wijngaarden-Cremers et al., 2014 for a review). This pattern emerges at least in part because RRBs are common in healthy preschoolers (Kim & Lord, 2012) and sex differences only become more starkly apparent in later childhood when RRBs are more atypical and impairing (South, Ozonoff, & McMahon, 2005).

After eliminating RRB items, I first examined item factor loadings in the one published factor analysis of the ADOS-2 Toddler Module (Guthrie et al., 2013) and excluded all items with a loading of less than 0.59 on the Social Communication and Social Interaction factor. I then chose to include only items that tapped into social behaviors that the research literature suggests are influenced, at least in part, by gendered socialization processes. I next examined each remaining item and included items that clinicians scored based on the child's behavior across the full testing session rather than on a single activity within the ADOS-2 Toddler Module. Finally, to reduce reliance on clinician subjectivity, I chose items that clinicians scored as a result of direct behavioral observations. After these four steps, eight items remained: B1. Unusual Eye Contact, B4. Facial Expressions Directed to Others, B5. Integration of Gaze and Other Behaviors During Social Overtures, B6. Shared Enjoyment in Interaction. B9. Requesting, B15. Quality of Social Overtures, B16b.Amount of Social Overtures: Parent/Caregiver, and C2. Imagination/Creativity.

I hypothesized that the selected ADOS-2 Toddler Module items would not comparably assess the latent variable of social communication in boys and girls. See Figure 1 for a visual representation of the hypothesized model. Support for this hypothesis would have several implications. First, it would suggest that current diagnostic items may be biased to more frequently identify toddler boys than girls as meeting ASD criteria. If selected social communication items were to differentially function for boys and girls, then findings would point to the possibility that algorithms should be modified for boys and girls to mitigate the potential for girls to receive scores that are lower than their true scores. Second, it would suggest that boys and girls may have different ASD phenotypes. Absence of support for this hypothesis would suggest that the ADOS-2 Toddler Module assesses ASD symptoms of social communication comparably across toddler boys and girls and that future research should focus on other possible explanations for the sex differences in diagnostic rates of ASD as assessed by the ADOS-2 Toddler Module.

The present study focused on a clinical sample of toddlers, because it is during this developmental period that children first regularly demonstrate many social and recreational behaviors that may be influenced by gendered socialization practices (e.g., Dunn, 1994; Wellman, Cross, & Watson, 2001). I used a clinical sample rather than a research sample. While the use of a clinical sample introduced confounds that threatened the internal validity of results, the clinical sample offered external validity; thus, I can more confidently generalize results to other community or clinical samples. An additional limitation of the current sample was that all of the toddlers were referred for an evaluation by a pediatrician or other provider due to suspected ASD. Thus, the fact that the sample includes more boys than girls could at least in part reflect referral biases.

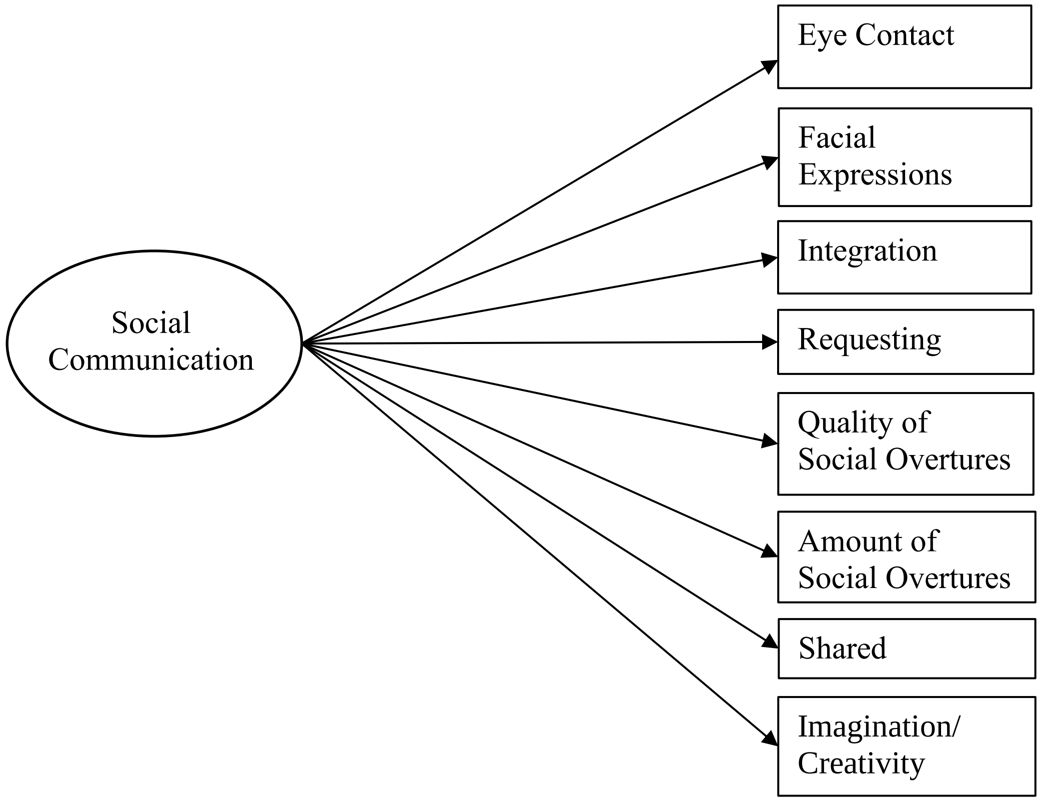

Figure 1 Schematic model of ADOS-2 Toddler Module items hypothesized to load on a Social Communication factor.

2 METHODS

2.1 Participants

The data analyzed in the present study were from a retrospective chart review conducted at a children's autism center in the southeastern United States. Children in the sample presented to the center's outpatient clinic for an ASD assessment between November 2017 and April 2019 following referral by their pediatricians. All referred children who were administered an ADOS-2 Toddler Module were included in the present sample.

I examined demographic variables for the sample (n = 315). Of the sample, 71.7% was male (n = 225). Children ranged in age from 14 to 30 months of age (M = 25 months; SD = 3.72 months). Within the sample, parents identified 156 children as White and the remainder as Black (n = 102; 49.5%), mixed race (n = 26; 8.3%), Asian/Middle Eastern (n = 12; 3.8%), or American Indian/Alaskan Native (n = 1; .03%). The majority of the sample was Not Hispanic/Latino (n = 256; 81.2%). Finally, 225 children (71.4%) had Medicaid as their primary insurance carrier and the rest of the sample had private insurance (n = 85; 27%). I identified those insured via Medicaid as low SES; those with any other insurance carriers were identified as high SES, consistent with prior research (Bach et al., 2002). The majority of the children within the sample received an ASD diagnosis (n = 250; 79%); of these children, 182 were boys (73%%) and 68 were girls (27%). A few children (n = 11; 3.5%) received a provisional ASD diagnosis (5 girls, 6 boys), meaning that he or she showed ASD characteristics, but not enough to meet criteria at the time of evaluation, and follow-up was recommended in a year. Within the remainder of the sample, 15 children (3 girls, 12 boys) received no diagnosis and the others received various other diagnoses, including Global Developmental Delay (n = 13), Disruptive

Behavior Disorder (n = 1), Language Disorder (n = 21), Speech-Sound Disorder (n = 1), and Unspecified or Other Neurodevelopmental Disorder (n = 4).

2.2 Procedure

A university/hospital Institutional Review Board approved all study procedures for this retrospective chart review. Caregivers provided consent for their child's data to be used in retrospective chart review studies. Children with suspected ASD and their caregivers presented to the autism center for an initial hour-long diagnostic interview. Those children whose symptoms warranted further evaluation for ASD returned with their caregivers for a three-hour autism assessment, during which caregivers and youth completed interviews, questionnaires, and psychological testing.

2.3 Measures

2.3.1 Family demographics

Prior to attending the diagnostic evaluation, caregivers of referred children completed the autism center's unpublished parent child questionnaire to provide demographic information. On this measure, they reported the referred child's sex, date of birth, and race/ethnicity. Additionally, families reported their insurance carriers; this information served in the present study as a proxy for socioeconomic status (SES).

2.3.2 Autism symptoms

The Autism Diagnostic Observation Schedule, 2nd edition Toddler Module (ADOS-2; Lord et al., 2012; Luyster et al., 2009) is a standardized, semi-structured observational assessment of autism symptoms; it provides a measure of social interaction, communication, play, and RRBs. The ADOS-2 Toddler Module, widely considered the gold standard ASD diagnostic measure (Falkmer et al., 2013), consists of five modules, each tailored to children of different ages and levels of language functioning. The Toddler Module (Luyster et al., 2009) was designed for trained clinicians to use with children who are between the ages of 12 and 30 months and who range from nonverbal to able to use simple two-word phrases. Psychologists trained to research reliability administered and scored the ADOS-2 Toddler Modules that I used in the present study (Cohen's kappa coefficient ≥ 0.80).

During the administration of the ADOS-2 Toddler Module, the examiner presents developmentally appropriate toys and activities in the context of a loosely structured testing session. Throughout the administration, the examiner documents behavioral observations, which provide the basis for determining scores. The Toddler Module comprises 11 activities, each of which consists of a hierarchy of social "presses," or planned social occasions, in which a range of social initiations and/or responses may occur (Lord et al., 1989). The 41 items that make up the ADOS-2 Toddler Module fall into five categories (Language and Communication, Reciprocal Social Interaction, Play, Stereotyped Behaviors and Restricted Interests, and Other Behaviors). Each item has between four and six response choices (e.g., 0 = typical behavior, 1 = some typical behavior, some atypical behavior, 2 = mostly atypical behavior, 3 = all atypical behavior, 7/8 = not applicable). The items included in the present study were: B1. Unusual Eye Contact, B4. Facial Expressions Directed to Others, B5. Integration of Gaze and Other Behaviors During Social Overtures, B6. Shared Enjoyment in Interaction. B9. Requesting, B15. Quality of Social Overtures, B16b. Amount of Social Overtures: Parent/Caregiver, and C2. Imagination/Creativity.

The ADOS-2 Toddler Module has adequate to good reliability and validity. In one study, inter-item correlations ranged from .18 to .82; Cronbach's alpha was .90 for the Social Affect domain and .50 for the RRB domain (Luyster et al., 2009). Test-retest reliability over two months ranged from .75 to .86, and the intraclass correlations for inter-rater reliability for the algorithms in the norming sample ranged from .74 to .99. The percentages of exact agreement on individual items level were all above 74% (Luyster et al., 2009). Regarding validity, the items in each domain correlated more strongly to each other than to items in other domains (Lord et al., 2012). Algorithms used to calculate a total score were found to discriminate between ASD and non-spectrum children with sensitivity (values range from .81 - .91) and specificity (values range from .83 to .94; Luyster et al., 2009). Total ADOS-2 Toddler Module scores have been found to correlate with verbal developmental functioning scores, measures of auditory comprehension and expressive language, and externalizing behavior problems, such that more severe autism symptoms were related to more severe other problems (Esler et al., 2015; Hedley et al., 2016).

2.3.3 Developmental functioning

The Mullen Scales of Early Learning (MSEL; Mullen, 1995) is a developmental assessment battery administered by trained clinicians and designed for use with children aged 0 to 68 months. For the present study, research reliable psychologists administered and scored the instrument; however, not all children received the full measure. The MSEL yields an Early Learning Composite score, which aggregates scores from five scales (Gross Motor, Visual Reception, Fine Motor, Receptive Language, and Expressive Language). However, the present study used only the Visual Reception scale as a measure of nonverbal developmental functioning (e.g., "Discriminates forms on formboard"). Response choices for each item comprise two to five possible choices, ranging from 0 to 1, 2, 3, or 4. A score of 0 represents the absence of a skill and higher scores indicate a demonstration of the skill. Standard T-scores range from 20 to 80; children who perform below this range receive a score of "<20". Thus, I also used age-equivalent scores to profile developmental functioning. I used mean difference testing to examine sex differences across nonverbal developmental functioning levels and presented descriptive data to further illustrate the sample.

The MSEL was standardized on a nationally representative sample and has acceptable to good psychometric properties (Mullen, 1995). Internal consistency ranges from .83 to .95 (Dumont, Cruse, Alfonso, & Levine, 2000) and test-retest and interrater reliability are good (Mullen, 1995). In terms of validity, scores on the MSEL have been found to correlate with scores on similar measures (i.e., Bayley Scales of Infant Development; Differential Abilities Scale; Infant-Toddler Social Emotional Assessment; Bishop, Guthrie, Coffing, & Lord, 2011; Carter, Briggs-Gowan, Jones, & Little, 2003).

2.4 Data Analytic Plan

I used a structural equation modeling (SEM) approach to differential item functioning (DIF) that relied on confirmatory factor analysis (CFA) methods. I examined whether the ADOS-2 Toddler Module assessed the latent variable of social communication, with indicators comprising eight selected items, with comparable effectiveness in toddler boys and girls referred for an ASD evaluation.

I considered several statistical approaches when deciding on a strategy to test the research question. Although mean difference testing and multivariate ANCOVA both permit examination of group differences, I determined that SEM was a better approach for the present study because it allowed me to make comparisons in the context of a latent variable measurement model, which adjusts for measurement errors and allows the separation of item properties from person performance (Embretson, 1996).

Different approaches to CFA with covariates also exist; these include multiple group and Multiple Indicators Multiple Causes (MIMIC) models (Brown, 2015). Both of these approaches have strengths. The MIMIC model is a graded response model with covariates that analyzes a single covariance matrix, which includes indicators and dummy codes for group membership. Essentially, a MIMIC model is a multivariate logistic regression for ordered response variables with latent variables. In this model, only latent factors and indicators are regressed onto a dummy-coded covariate. MIMIC models are parsimonious (i.e., each consists of a single matrix and few freely estimated parameters). The MIMIC model also has limitations. For example, it only examines indicator intercepts and factor means, so it assumes that all other measurement and structural parameters are the same across all levels of the covariate. MIMIC models also assume that the covariate is free of measurement error, although this assumption is potentially problematic when covariates are measured unreliably.

Multiple groups CFA analyzes data from discrete groups determined by the chosen covariate (i.e., two separate variance-covariance input matrices; one for boys and one for girls) and places constraints on the same parameters in both groups to examine measurement invariance and population heterogeneity. The main advantage of this approach is that is allows for examination of all aspects of measurement invariance and population heterogeneity, including factor loadings, intercepts, residual variances, factor variances, factor covariances, and latent means. It requires, however, a larger sample than does the MIMIC approach to have enough power to detect invariance, as the data are divided into discrete groups according to the covariate. Given my reasonable sample size (n = 315), I used the multiple group CFA approach.

Before conducting analyses, I first evaluated the extent of missing data and characterized missing data in terms of its relationships to the other study variables. Missing data can be characterized in one of three ways (Rubin, 1976). Ideally data are missing completely at random (MCAR), which means that they are missing for reasons unrelated to other study variables. For example, a clinician might have skipped an item inadvertently for a single participant. Data can also be missing at random (MAR) if they are absent for a predictable reason related to study variables. Girls, for instance, might be more likely than boys to refuse to complete a given item. Finally, data can be missing not at random (MNAR). In this case, data are missing as is a direct result of other study variables; children with the weakest social communication skills might be unable to complete particular items.

Examination of the data revealed that only 0.01% of data were missing. I could find no evidence that data were missing as a direct result of a study variable, nor were missing data predictable; thus, I determined that the missing points were MCAR. I used the full information maximum likelihood (FIML) assumption to address MCAR data (Little et al., 2014). FIML methods attend to missing data and estimate parameters and standard error in a single step (Graham, 2009) by fitting the model to the values that are not missing rather than estimating the missing values (Widaman, 2006).

I then screened data to ensure normality and absence of multivariate outliers. Next, I tested and compared a series of four models from least stringent equality constraints (i.e., parameters were held equal in value to each other) to most stringent equality constraints (i.e., more parameters were held equal to each other). I ran two sets of multiple group confirmatory factor analyses (MGCFA)—a categorical MGCFA and a linear CFA. For the first, I characterized the indicators as ordinal variables and for the second, I characterized them as continuous variables. While MGCFA and CFA are similar, there are a few key differences between them that I highlight throughout my description of the analyses. One key difference is that I used a maximum likelihood (ML) estimator and theta parameterization for continuous indicators and I used a weighted least square means and variance adjusted (WLSMV) estimator for ordinal items and theta parameterization in Mplus software (Version 8; Muthén & Muthén, 2005).

At step one, I tested baseline models for configural invariance (or equal form). Configural invariance was evident if the pattern of loadings of latent variable on the eight proposed indicators was similar in boys and girls. In the baseline model, all parameters (loadings, intercepts, and variances) were allowed to vary. In linear MGCFA, the parameters that are permitted to vary are the factor loadings, the item intercepts, and the item variances. In categorical MGCFA, the factor loadings and the item thresholds are permitted to vary. When the indicators are categorical or ordinal, the indicators do not have intercepts, but instead have thresholds above or below which the value of the latent factor falls into different groups (e.g., typical or atypical).

To evaluate these models, I examined X^2, comparative fit index (CFI) values, and root mean square error of approximation (RMSEA) values. CFI values > 0.90 indicated adequate fit, > 0.92 indicated good fit and > 0.95 indicated excellent fit (Hu & Bentler, 1999; Marsh, Hau, & Wen, 2004). RMSEA values > 0.10 indicated marginal fit, values > 0.08 and < 0.05 indicated adequate fit, and values < 0.05 indicated good fit (Browne & Cudeck, 1993; Marsh et al., 2004).

Second, I looked for strain in each solution and evaluated whether all freely estimated factor loadings were statistically significant ($p < .05$) and salient. If this model indicated good fit, I proceeded with the multiple groups CFA. To identify areas of strain, I examined standardized residuals (absolute values greater than 2.58 are problematic; Brown, 2015) and modification indices (values greater than 3.84 are problematic; Brown, 2015). Modification indices are estimates of how the overall fit of the model (i.e., X^2 value) would change if a certain parameter were added. Modification indices point to areas of strain in the model where there may be additional covariance between indicators or between indicators and the latent factor beyond what is specified in the model. If this model indicated good fit, this meant that there was configural invariance and the loadings of the indicators on the latent factor were the same for boys and girls. If this model did not fit, I planned to explore alternatives, such as dropping items associated with areas of strain, to improve the fit.

At step two, I tested for measurement invariance. Measurement invariance would indicate that the magnitudes of the factor loadings and item thresholds were similar across boys and girls. Both factor loadings and item thresholds were constrained to be equal (i.e., not permitted to vary) for boys and girls at this step. To determine if there was measurement invariance, I compared the fit of this model to the fit of the baseline (configural invariance) model by examining the change in overall model fit (i.e., change in X^2 and CFI). If the measurement invariance model indicated a lack of degradation in fit as compared to the configural invariance model (i.e., measurement invariance exists), the change in X^2 was statistically significant and the change in CFI was $< .01$.

When testing linear models, this step is broken into two parts. In the first part, I tested for weak invariance (or metric invariance or test of equal factor loadings). Metric invariance would indicate that the magnitudes of the loadings were similar across boys and girls. Factor loadings were constrained to be equal (i.e., not permitted to vary) for boys and girls in this step. To determine if there was metric invariance, I compared the fit of this model to the fit of the baseline (configural invariance) model by examining the change in overall model fit (i.e., change in X^2, CFI, and information criteria (i.e., Akaike information criteria [AIC] and sample-size adjusted Bayesian information criteria [BIC]). If the weak invariance model indicated a lack of degradation in fit as compared to the configural invariance model (i.e., weak invariance exists), the change in X^2 was statistically significant, the change in CFI was <.01, and the AIC and BIC decreased.

At the second part of this step for linear models, I tested for strong invariance (or scalar invariance). Scalar invariance is when the item intercepts are similar across groups. I did not permit factor loadings or item intercepts to vary in this model. To determine if there was scalar invariance, I examined the change in X^2, CFI, AIC, and BIC to determine whether the strong invariance model was not substantially worse than the weak invariance model and that strong invariance existed.

There is an optional step four, termed strict or residual invariance, aimed at determining whether residual variance varies across models. Some statisticians consider this step to be necessary, but others do not, given that residual variances include random error that is not expected to be the same for different models. This step is not typically completed for categorical models (Brown, 2016). For linear models at step four, I tested for strict invariance; in this model. I did not allow factor loadings, item intercepts, or residual variances to vary. To determine if there was strict invariance, I examined the change in X^2, CFI, AIC, and BIC to determine whether the strict invariance model was not substantially worse than the strong invariance model and that strict invariance existed.

Lastly, I conducted tests of population heterogeneity to compare models for boys and girls when two aspects of the latent variable were held equal across models. First, I held the latent variances equal, and then held the latent means equal. This final step can be done for both linear and categorical models. The purpose of these tests is to examine whether the latent factor of impaired social communication can be meaningfully compared across boys and girls.

3 RESULTS

3.1 Preliminary Analyses

Descriptive statistics for ADOS-2 items (Table 1) and Mullen visual reception scores (Table 2) are presented below. I included Mullen visual reception scores, which were available for 299 of the 315 participants, to help characterize the range of cognitive and motor functioning within the sample. Scores were reported in T-scores (*M* = 50, *SD* = 10) and ranged from the very low to average ranges of functioning. Scores were concentrated at the lower functioning end of the spectrum (*M* = 34.20; *SD* = 10.57), suggesting that most participants' functioning was below expectations for their chronological age.

Table 1 Descriptive Statistics: ADOS-2 Item Response Frequencies by Sex

ADOS-2 Tod Mod item	Responses							
	0		1		2		3	
	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
B1. Unusual Eye Contact	19	10	62	26	120	42	25	11
B4. Facial Expressions Directed to Others	16	8	84	32	101	37	25	12
B5. Integration of Gaze and Other Behaviors	26	10	71	34	106	35	23	10
B6. Shared Enjoyment in Interaction	81	36	75	20	51	24	19	9
B9. Requesting	50	24	34	14	127	41	15	10
B15. Quality of Social Overtures	29	6	80	38	96	39	21	6
B16b. Amount of Social Overtures/Maintenance of Attention: Parent/Caregiver	29	0	66	36	103	37	28	15
C2. Imagination/Creativity	11	10	62	21	48	17	104	41

I also examined sex differences on other demographic study variables. Independent *t*-tests revealed that no study variables significantly differed by sex. Results from non-parametric tests of sex differences in the ADOS-2 items revealed no significant sex differences (see Table 2).

Table 2 Study Variable Descriptive Statistics and Sex Differences

	Boys (n = 226)		Girls (n = 89)			
	M	SD	M	SD	*t*	Mann-Whitney *U*
Age	25.04	3.59	24.83	4.05	0.45	
Mullen: Visual Reception T-scores	34.55	10.81	33.31	9.94	0.92	
Mullen: Visual Reception Age equivalent (months)	18.98	4.78	18.20	4.66	1.30	
B1. Unusual Eye Contact						9665.50
B4. Facial Expressions Directed to Others						10050.50
B5. Integration of Gaze and Other Behaviors during Social Overtures						9616.00
B6. Shared Enjoyment in Interaction						9981.50
B9. Requesting						9770.50
B15. Quality of Social Overtures						9989.00
B16b. Amount of Social Overtures/Maintenance of Attention to Parent/Caregiver						9483.00
C2. Imagination/Creativity						9679.00

Note. $^*p < .05$

3.2 Single Group Analyses

To test the primary hypothesis, that eight selected social communication items in the ADOS-2 Toddler Module differentially assessed the latent social communication phenotype for girls and for boys, I first examined baseline model fit separately for each gender, allowing all parameters to vary. In the boys-only model, two data points (out of a total of 1808 data points) were missing, and in the girls-only model, there was one missing data point. I used a weighted least square means and variance adjusted (WLSMV) estimator for ordinal items and theta parameterization. To evaluate model fit, I examined X^2, comparative fit index (CFI), Tucker Lewis Index (TLI), and root mean square error of approximation (RMSEA) values.

3.2.1 Boys only

In the model with only boys, overall-fit statistics for the single-group solution indicated good model fit, $X^2(20) = 25.57$, $p = .12$, CFI = 1.00, TLI = 1.00, RMSEA [90% CI 0.00 – 0.08] = 0.04 (see Table 3). All unstandardized factor loadings were statistically significant and ranged in value from 0.74 to 1.05 (see Figure 2). Modification indices (MI) greater than 3.84 (Brown, 2015) were used to detect areas of strain in the model. There was no evidence of strain. When I tested the linear version of this model, it also indicated good overall fit (see Table 4).

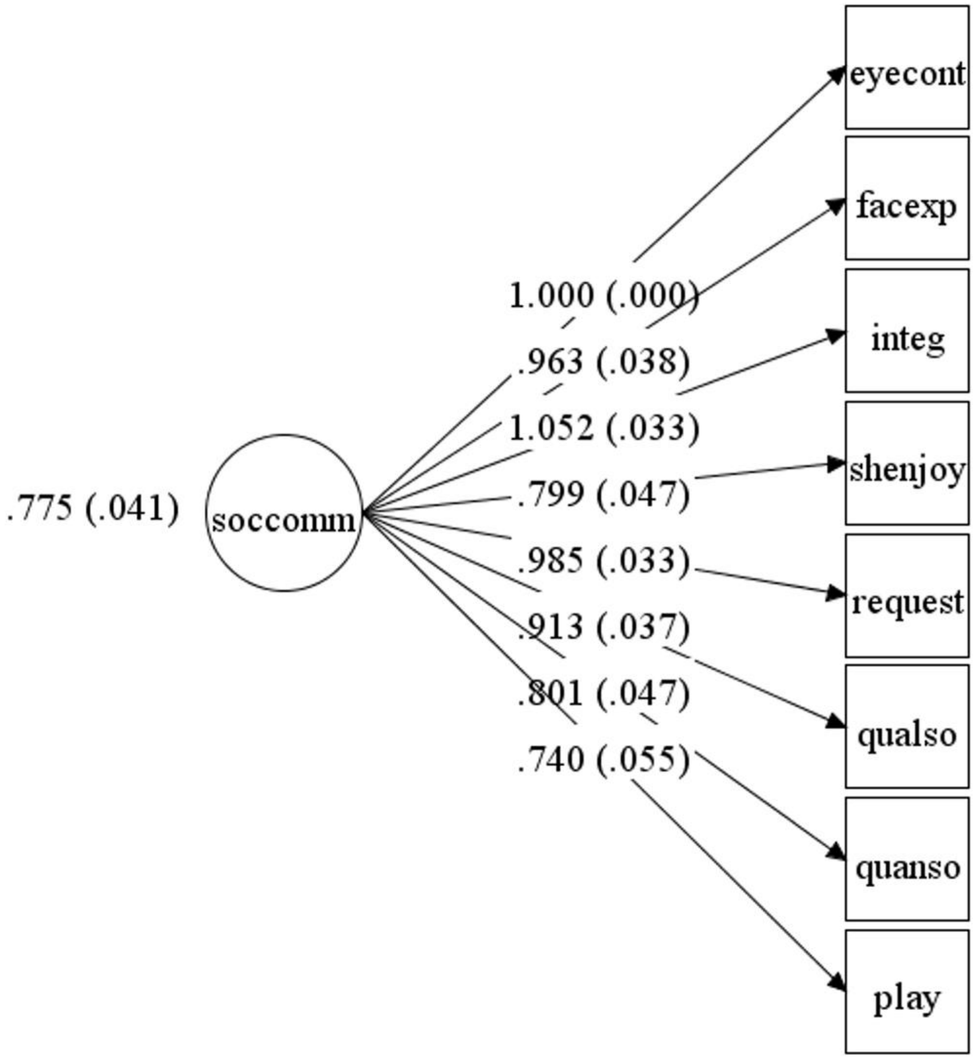

Figure 2 Unstandardized boys single group solution

Notes. *soccomm =latent social communication factor, eyecont = ADOS-2 B1 Unusual Eye Contact, facexp = ADOS-2 B5 Facial Expressions Directed to Others, integ = ADOS-2 B5 Integration of Gaze with Other Behaviors During Social Overtures, shenjoy = ADOS-2 B6 Shared Enjoyment in Interaction, request = ADOS-2 B9 Requesting, qualso = ADOS-2 B15 Quality of Social Overtures, quanso = ADOS-2 B16b Amount of Social Overtures/Maintenance of Attention: PARENT/CAREGIVER, play = ADOS-2 C2 Imagination/Creativity*

3.2.2 Girls only

In the model with only girls, overall-fit statistics for the single-group solution indicated good model fit, $X^2(20) = 16.18$, $p = .71$, CFI = 1.00, TLI = 1.00, RMSEA [0.00 – 0.07] = 0.00, (see Table 3). All unstandardized factor loadings were statistically significant and ranged in value from 0.63 to 1.08 (see Figure 3). For this model, MI did not indicate any areas of strain. Good overall fit was also evident for the linear girls-only model (see Table 4).

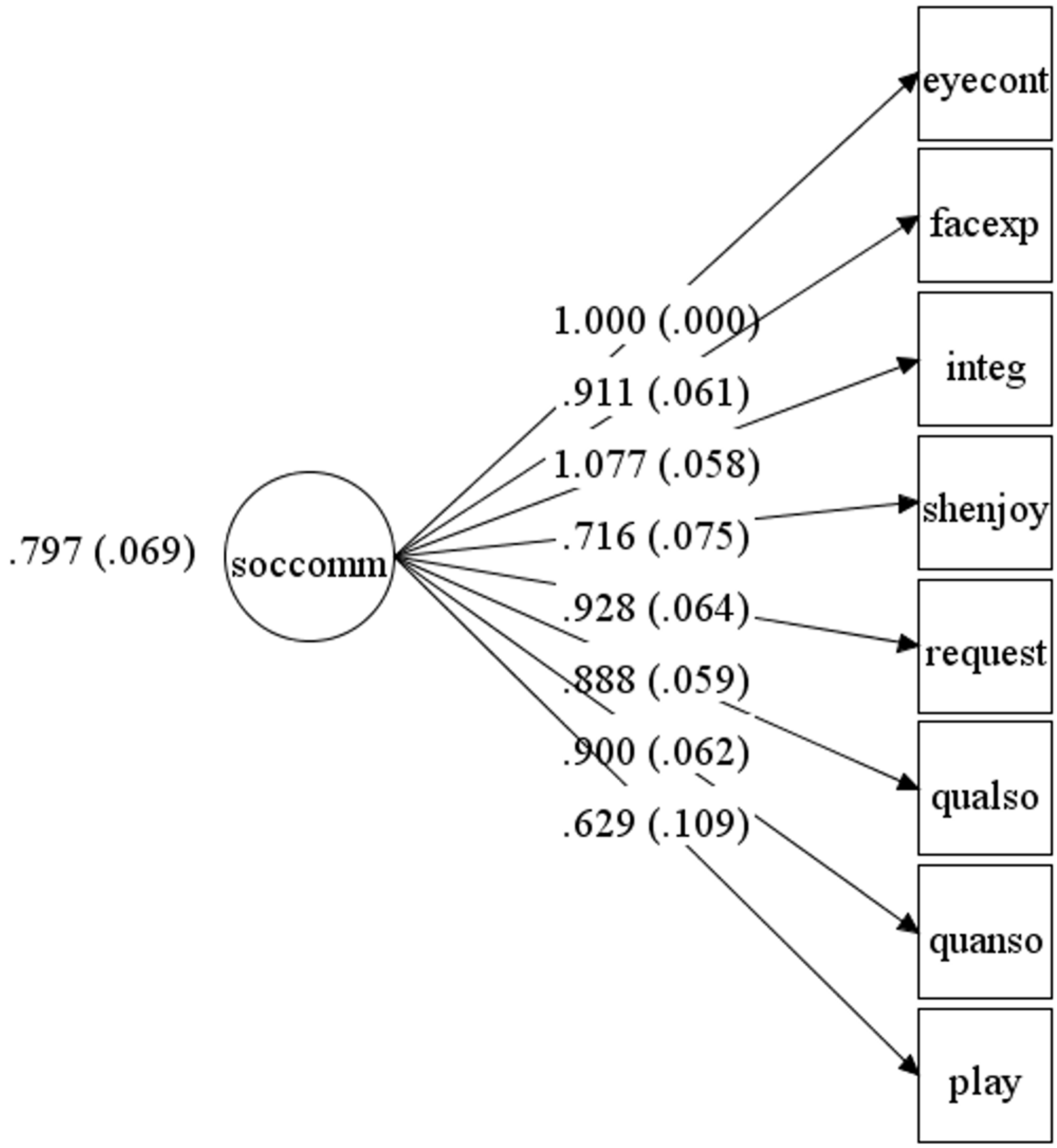

Figure 3 Unstandardized girls single group solution

Notes. *soccomm =latent social communication factor, eyecont = ADOS-2 B1 Unusual Eye Contact, facexp = ADOS-2 B5 Facial Expressions Directed to Others, integ = ADOS-2 B5 Integration of Gaze with Other Behaviors During Social Overtures, shenjoy = ADOS-2 B6 Shared Enjoyment in Interaction, request = ADOS-2 B9 Requesting, qualso = ADOS-2 B15 Quality of Social Overtures, quanso = ADOS-2 B16b Amount of Social Overtures/Maintenance of Attention: PARENT/CAREGIVER, play = ADOS-2 C2 Imagination/Creativity*

3.3 Measurement Invariance Analyses

The single-group solutions indicated good overall model fit for both boys and girls. Thus, I proceeded with the multiple groups CFA to examine measurement invariance of the models across boys and girls, proceeding through a series of models that progressed from the least constrained (i.e., fewer parameters held equal to each other) to more constrained (i.e., more parameters held equal to each other).

At step one of my tests for measurement invariance across groups, I conducted the analysis of equal form (also called the test of configural invariance), which tests the least restrictive model comparing the factor structure in boys and girls (i.e., no parameters are constrained to be equal to each other). When the model was tested with categorical MGCFA, item seven (Amount of Social Overtures/Maintenance of Attention: Parent Caregiver) was excluded because there were no "0" responses for girls (i.e., no clinicians rated girls as directing a typical number of social overtures to a parent or caregivers). With seven items, the categorical model indicated good overall fit, $X^2(42) = 34.98$, $p = 0.77$, CFI = 1.00, TLI = 1.00, RMSEA [0.00 – 0.04] = 0.00; with no significant areas of strain (see Table 4).

For boys and girls, all unstandardized factor loadings were statistically significant and ranged in value from 0.65 to 1.06. Loadings were identical to those in the single group solutions. Factor loading estimates revealed that all indicators were significantly related to the latent variable for boys and girls (range of R^2 = 0.34 - 0.87). MIs indicated no areas of strain in the model. Given that this model had good overall fit, I proceeded with the next model. There was also good overall fit for the continuous or linear version of this model (see Table 4).

In the categorical model, the next step compared the boy and girl models while holding factor loadings and item thresholds equal across boys and girls. There was good overall model fit, $X^2(41) = 32.74$, $p = 0.82$, CFI = 1.00, TLI = 1.00, RMSEA [0.00 – 0.04] = 0.00, and no significant areas of strain were evident. For boys and girls, all unstandardized factor loadings were statistically significant and ranged in value from 0.74 to 1.05. MIs indicated no areas of strain in the model (see Table 3).

To compare the fit of this model (i.e., the model with factor loadings and item thresholds are all held equal to each other) to the fit of the least restrictive model (i.e., the model where all parameters were permitted to vary), I examined the magnitude of changes in X^2. The change in X^2 was not significant, $X^2_{\text{diff}}(1) = 2.24$, $p >. 05$, which indicated that the more restrictive model did not significantly degrade fit. Since this model did not significantly degrade the model fit, I determined that it was the best fitting model or measurement invariance (see Figures 4 and 5) for categorical MGCFA (see Table 5 for factor loadings and item thresholds).

For linear models, this second step comprised into two components. I first conducted tests of equal factor loadings where the two models are compared to each other while factor loadings are held equal to each other across boys and girls. This test, also called test of weak, or metric, invariance, examines whether the magnitudes of the factor loadings are similar across groups. In other words, it allowed me to evaluate whether the ADOS-2 items have the same meaning and structure for boys and girls.

Overall fit statistics indicated good model fit for the linear model; $X^2(47) = 62.12$, $p = .07$, CFI = 0.99, TLI = 0.99, RMSEA [0.00 – 0.07] = 0.05; SRMR = 0.04; AIC = 5241.75; n-adj BIC = 5395.61 (see Table 4). For boys and girls, all standardized factor loadings were statistically significant and ranged from 0.52 to 0.90. R^2 values indicated that all factor loading estimates were significantly related to the latent variable (range of R^2= 0.27 - 0.81). MIs indicated a few areas of strain in the model. For boys, the MI for the latent variable by Quality of Social Overtures was 4.10, the MI for Requesting with Facial Expressions was 4.98, and the MI for Requesting with Integration was 5.43, and the MI for Imagination/Creativity with Integration was 6.12. For girls, the MI for the latent variable by Quality of Social Overtures was also 4.10, the MI for Quality of Social Overtures by Requesting was 4.14, and the MI for Quantity of Social Overtures with Integration was 5.74. The MI between the latent social communication factor and indicators suggest that there is extra covariance between those variables beyond what the model specifies.

To compare the fit of this linear model (i.e., the model with factor loadings are all held equal to each other) to the fit of the least restrictive model (i.e., the model where all parameters were permitted to vary), I then examined the magnitude of changes between models in information criteria (i.e., *n*-adjusted BIC and AIC) and in X^2. Both AIC and *n*-adjusted BIC increased, and the change in X^2 was not significant, $X^2_{diff}(6) = 6.05$, $p >. 05$, critical value of $X^2 = 12.59$, $df = 6$, $\alpha = .05$, which indicated that the equal factor loading model did not significantly degrade model fit. Since this model did not significantly degrade the model fit, I moved to the next step in the procedures that Brown (2015) outlined.

As a third step, I tested for strong invariance (or scalar invariance) to determine if the item intercepts were similar across groups. Factor loadings and item intercepts were both held equal to each other in this model. Overall fit statistics indicated good model fit; $X^2(54) = 67.62$, $p = .10$, CFI = 0.99, TLI = 1.00, RMSEA [0.00 – 0.07] = 0.04; SRMR = 0.04; AIC = 5233,25 *n*-adj BIC = 5252.99 (see Table 4). All standardized factor loadings were statistically significant and ranged from 0.52 to 0.90 and all factor loading estimates were significantly related to the latent variable (range of R^2 = 0.27 - 0.81). MIs indicated a few areas of strain in the model. For boys, the MI for the latent variable by Quality of Social Overtures was 4.12, the MI for

Requesting with Facial Expressions was 5.00, the MI for Requesting with Integration was 5.43, and the MI for Imagination/Creativity with Integration was 6.09. For girls, the MI for the latent variable by Quality of Social Overtures was also 4.12, the MI for Quality of Social Overtures with Requesting was 4.23, and the MI for Quantity of Social Overtures with Integration was 5.30.

When I compared the scalar invariance model (factor loadings and intercepts were constrained to be equal) to the weak invariance model (only factor loadings were constrained to be equal), AIC and *n*-adjusted BIC values decreased and the change in X^2 was not significant, $X^2_{diff}(7) = 5.50$, $p >. 05$ [critical value of $X^2 = 14.07$, $df = 7$, $\alpha = .05$. These results suggest that the fit of the model did not significantly degrade when I prevented the item intercepts from varying.

At the fourth step, which is optional for linear models, I tested for strict invariance—factor loadings, intercepts, and residuals are all held equal to each other across boys and girls. Overall fit statistics indicated good model fit; $X^2(62) = 81.18$, $p = .05$, CFI = 0.99, TLI = 0.99, RMSEA [0.00 – 0.07] = 0.04; SRMR = 0.06; AIC = 5230.73; *n*-adj BIC = 5245.83 (see Table 4). All standardized factor loadings were statistically significant and ranged from 0.55 to 0.89 and all factor loading estimates were significantly related to the latent variable (range of $R^2 = 0.31$ - 0.78). MIs indicated a few areas of strain in the model. For boys, the MI for the latent variable by Quality of Social Overtures was 4.60, the MI for Requesting with Integration was 4.82, and the MI for Imagination/Creativity with Integration was 5.24. For girls, the MI for the latent variable by Quality of Social Overtures was also 4.66, the MI for Quality of Social Overtures with Requesting was 4.23, the MI for Quantity of Social Overtures with Integration was 4.96, the

MI for Imagination/Creativity with Requesting was 4.03, and the MI for Imagination/Creativity with Quantity of Social Overtures was 4.72.

When I compared the strict invariance model (factor loadings, intercepts, and residuals were constrained to be equal) to the scalar invariance model (factor loadings and intercepts were constrained to be equal), AIC and *n*-adjusted BIC values increased and the change in X^2 was significant, $X^2_{\text{diff}}(8) = 13.48$, $p >. 05$, critical value of $X^2 = 15.51$, $df = 8$, $\alpha = .05$. These results suggest that the fit of the model did not significantly degrade when I prevented the item residual variances from varying.

3.4 Population Heterogeneity Analyses

Finally, I conducted tests of population heterogeneity for both the linear and categorical MGCFAs to determine if the latent factors can be meaningfully compared across boys and girls. For the linear models, overall fit was acceptable and for the categorical models, overall fit was good. When compared to the best fitting measurement invariance models (i.e., scalar invariance for the linear MGCFA and measurement invariance for the categorical MGCFA), there was no evidence of degradation in fit. See Table 3 for the categorical model fit and Table 4 for linear model fit.

Table 3 Categorical Measurement Invariance

	X^2	df	RMSEA [90% CI]	CFI	TLI	Comparison X^2_{diff}	Δdf	Notes
Single Group Solutions								
Boys	25.57	20	0.04 [0.00 – 0.08]	1.00	1.00			
Girls	16.18	20	0.00 [0.00 – 0.87]	1.00	1.00			
Boys (no item 7)	18.28	14	0.04 [0.00 – 0.08]	1.00	1.00			
Girls (no item 7)	9.68	14	0.00 [0.00-0.07]	1.00	1.01			
Measurement Invariance								
Configural Invariance	34.98	42	0.00 [0.00 – 0.04]	1.00	1.00			
Measurement Invariance	32.74	41	0.00 [0.00 – 0.04]	1.00	1.00	2.24	1	Compared to configural
Population Heterogeneity								
Latent Variances	38.20	48	0.00 [0.00 – 0.03)	1.00	1.00	5.46	7	Compared to measurement invariance
Latent Means	37.66	49	0.00 [0.00 – 0.03]	1.00	1.00	4.92	8	Compared to measurement invariance

Note. RMSEA=root mean square error of approximation, CI=confidence interval, CFI=comparative fit index; TLI=Tucker-Lewis index; $*p < .05$

Table 4 Linear Measurement Invariance

	X^2	df	RMSEA [90% CI]	SRMR	CFI	TLI	X^2_{diff}	Δdf	n-adj BIC	AIC	Notes
Single Group Solutions											
Boys	34.18*	20	0.06 [0.02 – 0.09]	0.03	0.99	0.98			3677.48	3677.48	
Girls	21.53	20	0.03 [0.00 – 0.10]	0.03	1.00	0.99			1555.85	1571.87	
Measurement Invariance											
Configural Invariance	56.07	41	0.05 [0.00 – 0.08]	0.03	0.99	0.98			5275.00	5247.70	
Metric Invariance	62.12	47	0.05 [0.00 – 0.07]	0.04	0.99	0.99	6.05	6	5395.61	5241.70	Compared to configural
Scalar Invariance	67.62	54	0.04 [0.00-0.07]	0.04	0.99	1.00	5.50	7	5252.99	5233.25	Compared to metric
Residual/Strict Invariance	81.10	62	0.04 [0.00-0.07]	0.06	0.99	0.99	13.48	8	5245.83	5230.73	Compared to scalar
Population Heterogeneity											
Latent Variances	102.42*	64	0.06 [0.03-0.08]	0.09	0.97	0.98	34.80	10	5261.99	5248.05	Compared to scalar
Latent Means	102.47*	65	0.06 [0.04-0.08]	0.09	0.97	0.98	34.85	11	5246.10	5259.46	Compared to scalar

Note. RMSEA=root mean square error of approximation, CI=confidence interval, SRMR=standardized root mean square residual, CFI=comparative fit index; TLI=Tucker-Lewis index; n-adj BIC=sample size adjusted Bayesian information criterion; AIC=Akaike information criterion; $*p < .05$

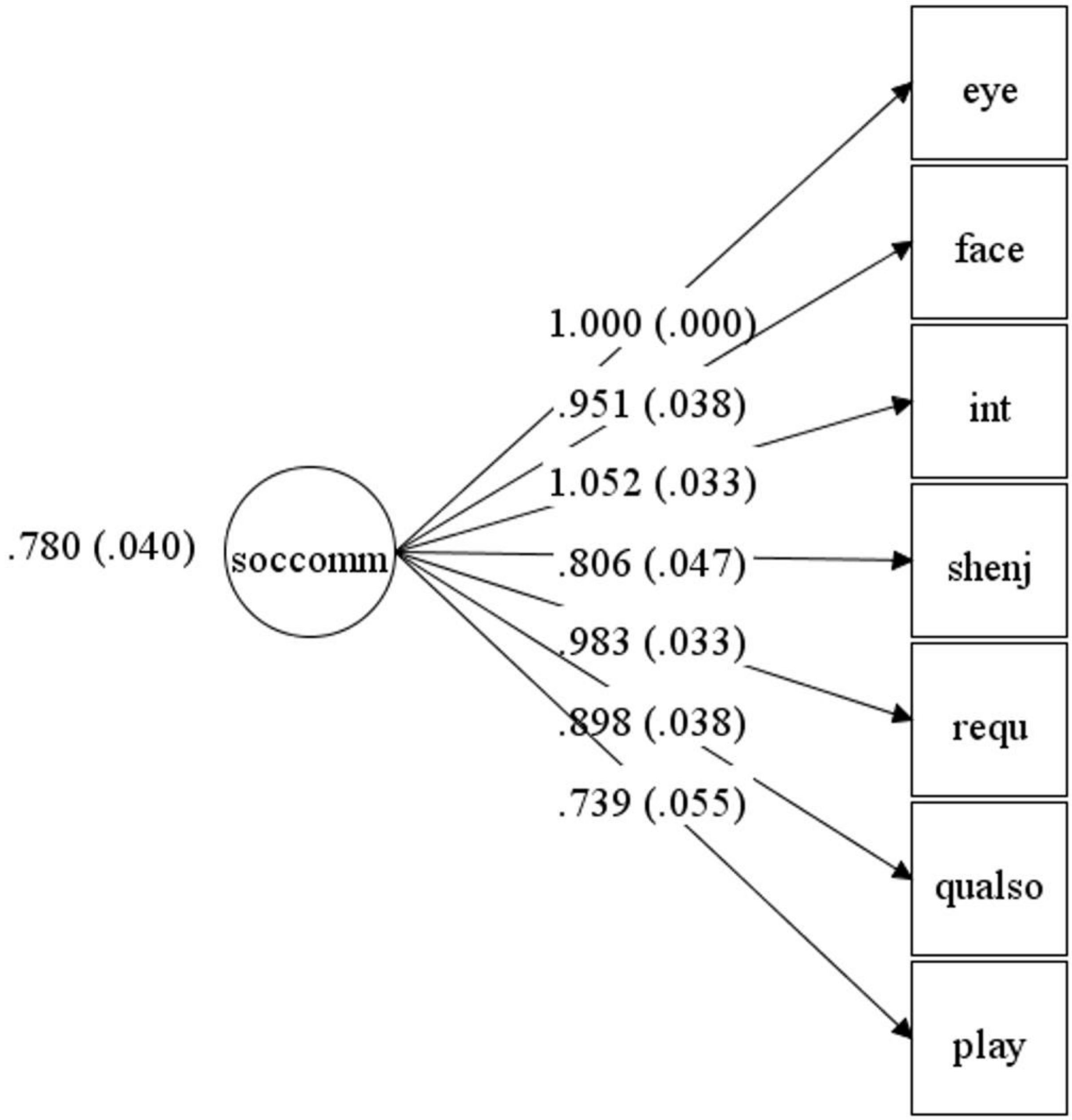

Figure 4 Boys unstandardized factor loadings for measurement invariance

Notes. *soccomm =latent social communication factor, eyecont = ADOS-2 B1 Unusual Eye Contact, facexp = ADOS-2 B5 Facial Expressions Directed to Others, integ = ADOS-2 B5 Integration of Gaze with Other Behaviors During Social Overtures, shenjoy = ADOS-2 B6 Shared Enjoyment in Interaction, request = ADOS-2 B9 Requesting, qualso = ADOS-2 B15 Quality of Social Overtures, quanso = ADOS-2 B16b Amount of Social Overtures/Maintenance of Attention: PARENT/CAREGIVER, play = ADOS-2 C2 Imagination/Creativity*

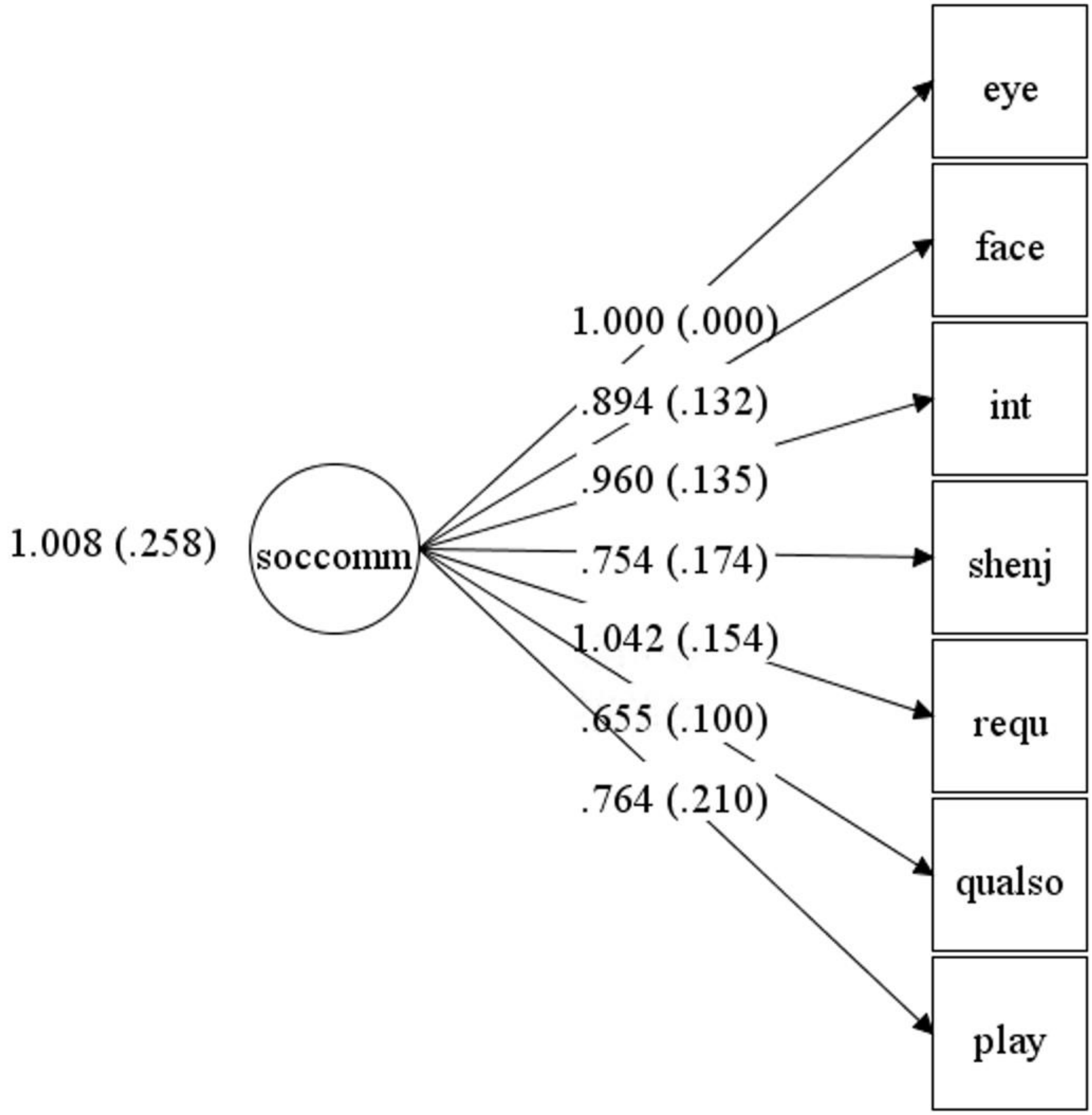

Figure 5 Girls unstandardized factor loadings for measurement invariance

Notes. *soccomm =latent social communication factor, eyecont = ADOS-2 B1 Unusual Eye Contact, facexp = ADOS-2 B5 Facial Expressions Directed to Others, integ = ADOS-2 B5 Integration of Gaze with Other Behaviors During Social Overtures, shenjoy = ADOS-2 B6 Shared Enjoyment in Interaction, request = ADOS-2 B9 Requesting, qualso = ADOS-2 B15 Quality of Social Overtures, quanso = ADOS-2 B16b Amount of Social Overtures/Maintenance of Attention: PARENT/CAREGIVER, play = ADOS-2 C2 Imagination/Creativity*

Table 5 Factor Loadings and Item Thresholds for Categorical Measurement Invariance Models

Relation/Variable	Unstandardized Estimate	Standard Error	Ratio	*p*	Standardized Estimate
Boys: Factor Loadings					
Social Communication BY					
Eye Contact	1.00	0.00	999.00	999.00	0.88
Facial Expressions	0.95	0.04	25.34	0.00	0.84
Integration of Gaze	1.05	0.03	31.47	0.00	0.93
Shared Enjoyment	0.81	0.05	17.30	0.00	0.71
Requesting	0.98	0.03	30.23	0.00	0.87
Quality of Social Overtures	0.90	0.04	25.46	0.00	0.79
Imagination/Creativity	0.74	0.06	13.52	0.00	0.65
Girls: Factor Loadings					
Social Communication BY					
Eye Contact	1.00	0.00	999.00	999.00	0.91
Facial Expressions	0.89	0.13	6.75	0.00	0.82
Integration of Gaze	0.96	0.14	7.08	0.00	0.94
Shared Enjoyment	0.75	0.17	4.34	0.00	0.64
Requesting	1.04	0.15	6.77	0.00	0.83
Quality of Social Overtures	0.66	0.10	6.55	0.00	0.80
Imagination/Creativity	0.76	0.21	3.65	0.00	0.59
Boys: Item Thresholds					
Eye Contact					
1	-1.38	0.12	-11.88	0.00	-1.38
2	-0.35	0.08	-4.23	0.00	-0.35
3	1.23	0.11	11.67	0.00	1.23
Facial Expressions					
1	-1.48	0.12	-12.20	0.00	-1.48
2	-0.15	0.08	-1.98	0.05	-0.15
3	1.21	0.11	11.56	0.00	1.21
Integration of Gaze					
1	-1.22	0.11	-11.46	0.00	-1.22
2	-0.14	0.08	-1.79	0.07	-0.14
3	1.26	0.11	11.60	0.00	1.26
Shared Enjoyment					
1	-0.35	0.08	-4.34	0.00	-0.35
2	0.47	0.08	5.68	0.00	0.47
3	1.40	0.12	11.97	0.00	1.40
Requesting					
1	-0.78	0.09	-8.52	0.00	-0.78
2	-0.31	0.08	-3.80	0.00	-0.31
3	1.50	0.13	11.88	0.00	1.50
Quality of Social Overtures					
1	-1.17	0.10	-11.56	0.00	-1.17
2	-0.04	0.07	-0.55	0.58	-0.04

3	1.29	0.11	11.93	0.00	1.29
Imagination/Creativity					
1	-1.64	0.14	-11.72	0.00	-1.64
2	-0.47	0.08	-5.77	0.00	-0.47
3	0.10	0.08	1.23	0.22	0.10
Girls: Item Thresholds					
Eye Contact					
1	-1.38	0.12	-11.88	0.00	-1.38
2	-0.35	0.08	-4.23	0.00	-0.35
3	1.23	0.11	11.67	0.00	1.23
Facial Expressions					
1	-1.48	0.12	-12.20	0.00	-1.48
2	-0.15	0.08	-1.90	0.06	-0.15
3	1.21	0.11	11.56	0.00	1.21
Integration of Gaze					
1	-1.22	0.11	-11.46	0.00	-1.22
2	-0.14	0.08	-1.79	0.07	-0.14
3	1.26	0.11	11.60	0.00	1.26
Shared Enjoyment					
1	-0.35	0.08	-4.34	0.00	-0.35
2	0.47	0.08	5.68	0.00	0.47
3	1.40	0.12	11.97	0.00	1.40
Requesting					
1	-0.78	0.09	-8.52	0.00	-0.78
2	-0.31	0.08	3.80	0.00	-0.31
3	1.50	0.13	11.88	0.00	1.50
Quality of Social Overtures					
1	-1.17	0.10	-11.56	0.00	-1.17
2	-0.04	0.07	0.55	0.58	-0.04
3	1.29	0.11	11.93	0.00	1.29
Imagination/Creativity					
1	-1.64	0.14	-11.72	0.00	-1.64
2	-0.47	0.08	-5.77	0.00	-0.47
3	0.10	0.08	1.23	0.22	0.10

4 CONCLUSIONS

ASD is more often diagnosed in boys than in girls (e.g., Shaw et al., 2020) and boys tend to receive diagnoses earlier than do girls (McCormick et al., 2020). Girls' delayed and relatively infrequent receipt of an ASD diagnosis may interfere with their access to early intervention. To examine the possibility that biased measurement at the level of individual assessment items accounts, at least in part for the gender differences, I examined whether eight ADOS-2 Toddler Module items function differently for boy and girl toddlers with suspected ASD. I focused in particular on eight ADOS-2 Toddler Module items that presumably measure social communication behavioral deficits common among individuals with ASD. I selected items that tap behaviors for which the research literature provides theoretical and empirical reasons to expect sex differences. I hypothesized that there would be sex differences at the item level, such that the magnitude of the relationships of the indicators to the social communication latent factor would differ across boys and girls.

Through comparing a series of increasingly restrictive models between boys and girls, my data indicate that, contrary to predictions, the ADOS-2 items related to the latent social communication factor almost identically for boys and girls. For both sexes, unstandardized factor loadings ranged from 0.64 (Imagination/Creativity) to 1.07 (Integration of Gaze and Other Behaviors). These unexpected findings merit discussion on at least two fronts. First, I focus on the observed factor structure and item loadings, with particular attention to the item—Imagination/Creativity —that loaded most weakly for both boys and girls. Second, I examine potential reasons that I may have failed to observe sex differences on the other items that showed consistently strong loadings for children of both sexes.

4.1 Factor Loadings and Item Conceptual Fit with the Social Communication Domain

The observed loading pattern is partly consistent with results from the one published factor analysis of ADOS-2 Toddler Module items in a mixed sample of boys and girls (Guthrie et al., 2013). Guthrie and colleagues (2013) found that the best fitting model for their data comprised two factors, one of which they labeled social communication. Seven of the eight items that I examined in the current study loaded onto Guthrie et al.'s (2013) social communication factor. Imagination/Creativity was the one exception; Guthrie and colleagues found that it did not load onto either factor. It is possible that this discrepancy reflects differences in scope between the present study and Guthrie et al. (2013). Whereas I selected items a priori that I expected both to load on a social communication factor and to be likely, for theoretically and empirically guided reasons, to function differently for boys and girls, Guthrie and colleagues included all ADOS-2 Toddler Module items in their models. Had I included a broader range of items, it is possible that Imagination/Creativity, which loaded the most weakly of the examined items on the Social Communication factor, might have been more closely associated with other latent constructs. Taken together, my findings and those of Guthrie et al. suggest, at a minimum, that how deficits in imagination/creative play fit into the underlying ASD structure remains unclear.

A lack of clarity about how deficits in imagination/creativity fit in ASD is longstanding and may reflect, at least in part, the fact that play is a complex entity that is difficult to define (Burghardt, 2011). To determine which activities constitute play, researchers consider the degree to which an individual engages in them flexibly, they engender positive affect, they involve a nonliteral stance, and they occur as a function of intrinsic motivation (Krasnor & Pepler, 1980). Moreover, the ADOS-2 Creativity/Imagination item focuses specifically on creative play, a subset of play that necessarily involves the creation of a non-real or imaginary context (Lillard, 1993). Regardless of whether children engage in creative play with other children or independently, this type of play is inherently social, in that participants are required to view things from perspectives other than their own.

Capturing play in a single item, as the ADOS-2 Toddler Module does, thus requires clinicians to integrate a great deal of information that may be difficult to extract from subjectively interpreted observations of behavior. The Imagination/Creativity item thus differs in notable ways from the other behaviors which the ADOS-2 Toddler Module assesses, such as directed eye contact and use of directed facial expressions, that I included in the present study. First, it requires that observers integrate several pieces of subjective information to yield a single score. In contrast, other items simply involve marking the presence or absence of objectively observed behaviors.

Second, whereas behaviors such as directed eye contact serve simple social communication functions, imagination/creativity requires that children integrate social communication with a capacity for other foundational skills such as imitation, turn-taking, and perspective-taking. Creative play skills emerge along a typical developmental trajectory that builds on itself, starting with parallel play, progressing to imitative play and simple turn-taking play with others, and culminating in fully creative play (Howes, 1980; Howes, 1988). Capturing where a child's play falls along this developmental trajectory is also complicated to do in the context of a single item.

Given the inherent complexity of creative play, it is also possible that clinicians administering the ADOS-2 are less reliable in their ratings of its presence or absence, as well as its quality, than they are in their ratings of other behaviors such as eye contact. Luyster and colleagues (2009), however, found that inter-rater reliability coefficients for all ADOS-2 Toddler Module items exceeded .71, which is considered good. This finding suggests that rating reliability may be less of an issue than conceptual complexity of play per se.

In recognition of these issues, the current diagnostic algorithm in the ADOS-2 Toddler Module excludes scores on the Imagination/Creativity item. Although the original ADOS algorithm included the Imagination/Creativity item (Lord et al., 1989), the authors removed it in subsequent revisions (e.g., ADOS-G, ADOS-2) because of concerns that it was too brief to allow clinicians to evaluate a child's imagination/creativity (Lord et al., 2000). We may thus genuinely lack the data that would help clarify how imaginative or creative play is impaired in the context of ASD and whether it might show differential impairment between boys and girls. Development of measures that permit efficient, but thorough, assessment of this set of complex behaviors may be necessary to clarify how deficits in creative/imaginative play fits within the greater ASD structure.

4.2 Potential Explanations of Why No Sex Differences Emerged for Other Items

The other ADOS-2 Toddle Module items that I examined fit more neatly into the conceptual construct of social communication. However, despite theoretical and empirical reasons from both the typically developing and the ASD literature that subtle sex differences exist (e.g., Malatesta & Haviland, 1982), I detected no evidence that they related to the latent social communication variable differently for boys and girls. There are several reasons sex differences may not have been evident for these social communication behaviors in the current sample at the item level.

First, the present study focused on children referred for a thorough ASD evaluation, typically by their primary care providers (PCPs), as is the case at many autism centers. Consequently, in light of evidence that PCPs and other healthcare providers more often refer males than females (e.g., Mandy et al., 2012), a referral bias may have affected results from the current study. Moreover, not only might a gender bias exist, but those girls who did get referred may have exhibited the most apparent indicators of developmental delay. PCPs may have overlooked girls with subtler symptoms.

A gender bias at this step could have contributed to the failure to detect sex differences in the present study. In particular, it could have meant that the current sample did not include enough girls with suspected ASD to permit the detection of subtle sex differences at the behavior level. Additionally, if such a bias were to reflect a tendency for referring providers to miss subtle signs of ASD in toddler girls, it would presumably result in different symptom distributions for male and female participants. For example, girls may be less frequently represented at the higher end of the spectrum (i.e., children with higher developmental functioning scores) than are boys. The current study lacks the data needed to examine this possibility of a gender bias at the level of the referring provider; further research into sex differences at the initial referral and identification of concerns about child behavior and functioning could help clarify whether referral thresholds may need to vary by sex in order to permit adequate sampling of girls.

Second, not only might referral biases have decreased the likelihood that sex differences in item performance would emerge, but the age range of the children under study—12 to 30 months—may also have been an issue. In particular, the more complex behaviors that the ADOS-2 examines may not evidence nuanced sex differences until children are past the toddler period. For example, school-age girls demonstrated better creative play skills than boys per parent report in one study (Beggiato et al., 2017) and better reciprocal conversation skills per parent and clinician report in another (Hiller et al., 2014).

These observed differences may reflect the fact that, as children age, not only do their social behaviors become more complex and sophisticated (e.g., Berry & O’Connor, 2010), but the sustained effects of gendered socialization may become more pronounced. Typically developing toddlers progress quickly to flexibly engaging in more complex and sophisticated interactions (Killen & Smetana, 1999; Killen & Turiel, 1991); for example, by integrating their social communication skills (e.g., gaze, gestures) with more sophisticated language (Conlon et al., 2019; Goddard et al., 2014). As their social communication skills develop, older children also face more nuanced social demands, which, in turn, contribute to the development of these increasingly complex and sophisticated social communication skills. Thus, the limited repertoire of social communication strategies that toddlers employ, as a function of, at least in part, their simpler behaviors and fewer social demands, may result in a narrower range of possible social communication behaviors, making individual differences more difficult to discern.

A third possible explanation as to why there was no evidence of sex differences in the current sample revolves around the social interaction elements that the ADOS-2 items evaluate. The items included in the current study assess superficial details of social interactions, which are typically stereotyped (e.g., making eye contact to initiate social interactions) rather than effective, nuanced social interactions that unfold over time. They thus may be less likely to yield sex differences than would items that focus on more complex social behaviors, but that are difficult to assess, especially in very young children. Assessing superficial social behaviors may be appropriate in evaluations of very young children, as superficial social interactions precede more complex ones (Guaralnick & Winehouse, 1984); however, learning how, and if it is even possible, to assess the flexible maintenance of toddlerhood social interactions may be beneficial for early ASD detection.

Research with older children and adults with ASD examining sex differences in the flexible maintenance of social interactions over time also suggests that this is an area to examine further (e.g., Dean et al., 2017; Tierney et al., 2015). Consistent with social camouflage theory, older girls and women with ASD often report understanding the initial mechanics of social interactions (e.g., smile, make eye contact, say hello), but then struggle to maintain and be flexible within those interactions (Milner et al., 2019; Tierney et al., 2016). The extent to which these findings generalize to toddlerhood is unclear.

In summary, findings from the present study suggest that the ADOS-2 Toddler Module measures seven social communication and play behaviors in similar ways for boys and girl toddlers with suspected ASD. The current study's lack of evidence of sex differences at the item level may be due to sampling factors (i.e., a referral bias), developmental factors (e.g., toddlerhood), and measurement factors (e.g., ADOS-2 Toddler Module items). Looking forward, examining sex differences at the symptom level in samples of older children and a broader range of children not just referred for an autism evaluation may be beneficial in addressing these concerns. For instance, item-level data from parent or clinician report of social behaviors of all children who present to a primary care clinic and not just those who display social difficulties as noted by PCPs would decrease the impact of referral biases. Additionally, further examination of how autism assessments measure the flexible maintenance of social interactions, not just the stereotyped aspects of social interactions, may help clarify if sex differences exist at different points during the interactions, especially in young children. Researchers have been developing creative ways to assess the maintenance of social interactions during this early developmental stage. For example, there is a body of emerging research that eye tracking to investigate what infants focus on and visually follow in social situations (e.g., Constantino et al., 2017).

4.3 Limitations

Although the current study made distinctive contributions to the literature examining sex differences in social communication behaviors characteristic of ASD, it has several limitations that warrant attention. First, the sampling strategy, which offers numerous advantages, also has potential drawbacks. Participants comprised a prospective clinical sample that encompassed only children whose PCPs referred them for an ASD evaluation at a clinic specializing in developmental disability evaluations. The benefit of using the sample in the present study is that it is likely similar to the clinical populations that present to other autism testing centers, which increases the chance that the results can be generalized. However, the present clinical sample may not be fully representative of the complete autism spectrum, as girls and children from non-majority racial/ethnic groups often do not present for an evaluation until after toddlerhood (McCormick et al., 2020; Smith et al., 2020). Additionally, some of the children included in the present study did not receive an ASD diagnosis (roughly 20% of the sample), so those children exhibited no social difficulties consistent with ASD[1].

In addition to possible issues associated with diagnostic heterogeneity, the sample may also have issues associated with participant ASD severity. In particular, the sample may have been skewed toward children with more severe symptoms, given that PCPs typically refer children for specialized evaluations who exhibit disruptive behavior concerns or language delays that warrant urgent attention (Sacrey et al., 2017). Although this sample captures toddlers with developmental functioning that spans the very low to average ranges, the sample may not fully represent the breadth of the autism spectrum, in that it includes a limited number of children with less severe symptoms, who are less likely to be flagged for an ASD evaluation by their PCPs.

The demographic makeup of the sample also warrants attention. Not surprisingly, there were more boys than girls in the sample, which could be a result of referral bias. However, this sample is unusual in that over half of the toddlers are non-White. Closer examination of the

[1] Exploratory multiple group CFAs conducted separately for participants who met ASD criteria and a broader group diagnosed with ASD/provisional ASD sample yielded no evidence of statistically significant sex differences.

racial composition of the sample reveals that the racial breakdown changes slightly with ASD severity. At the higher functioning end of the spectrum as indexed by Mullen visual reception scores, White toddlers are the best represented racial group (63%); at the lower end of the spectrum, the best represented racial group is Black/African American (47%). Minority children are more often diagnosed later than are those from majority groups, and those who are diagnosed earlier tend to have more severe behavioral and language concerns, perhaps because their issues attract attention more quickly (McCormick et al., 2020; Smith et al., 2020). While the current sample may offer a better representation of the general US population than many previous studies have, it is difficult to know how well it maps onto the true ASD population.

Second, in addition to sampling-related limitations, there are issues with the structure and administration of the ADOS-2 that warrant consideration. The ADOS-2 consists of a short observation session in a novel environment with a novel person. This short span of time may not allow observers to fully and accurately witness and document a child's typical range of behavior. Additionally, ADOS-2 scoring largely relies on clinical judgment. Despite training and reliability assurances, clinician bias exists and is inherent in the administration and scoring. This issue may be particularly salient for items such as creative play, which is a more complex behavior requiring the integration of several social skills (e.g., reciprocity).

Third, although the statistical approach that I took in the present study allowed me to examine subtle measurement differences and the relationships of particular items to a latent factor in multiple groups, it, too, has limitations. Sample sizes were not equal—multiple group solutions can be evaluated when the group sizes are different, however, variable group sizes may influence results (Brown, 2015). For example, the male sample contributes more cases to the overall chi-square values than the female sample, likely affecting overall fit indices and subsequent interpretation. Several other values are based on chi-square (e.g., CFI, modification indices) or are independently sensitive to sample size (e.g., standard errors, standardize residuals), so must be interpreted with caution.

Lastly, I based my hypotheses on a theoretical model that only considered socialization processes as potential contributors to sex differences. However, it is important to note that the factors involved in sex differences are many and varied; for instance, genetic factors (e.g., Acuna-Hidalgo et al., 2016; Nguyen et al., 2020; Snijders et al., 2015; Turner et al., 2019). Looking forward, research examining the interplay of genetic and social factors may be beneficial to further understanding the sex discrepancy in ASD.

4.4 Conclusions

Results of the current study of sex differences in performance on select ADOS-2 Toddler Module social communication items were inconsistent with hypotheses and yielded no evidence of sex differences at the item level. The main implication of this finding is that these ADOS-2 Toddler Module items are not inherently gender-biased in the current sample. This evidence suggests that the ways in which the ADOS-2 measures seven social communication behaviors in toddlers with suspected ASD are not major contributors to the sex differences observed in ASD.

This study was an important first step toward understanding sex differences in ASD by examining them at a granular, symptom level. As this emerging area of research continues to grow, it will be valuable to continue examining how boys and girls with ASD differ or are similar at the behavior level and the mechanisms (e.g., socialization, genetics, referral biases, racial/ethnic differences) that contribute to those differences or similarities. Further understanding of the factors that do contribute to sex differences will facilitate earlier diagnoses and access to intervention for girls.

REFERENCES

Acuna-Hidalgo-R., Veltman, J. A., & Hoischen, A. (2016). New Insights into the Generation and Role of De Novo Mutations in Health and Disease. Genome Biology, 17(241), 1-19.

Adams, S., Kuebli, J., Boyle, P. A., & Fivush, R. (1995). Gender differences in parent-child conversations about past emotions: A longitudinal investigation. *Sex Roles*, *33*(5-6), 309-323.

Alexander, G. M., Wilcox, T., & Woods, R. (2009). Sex differences in infants' visual interest in toys. *Archives of Sexual Behavior*, *38*(3), 427-433.

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

Bach, P. B., Guadagnoli, E., Schrag, D., Schussler, N., & Warren, J. L. (2002). Patient demographic and socioeconomic characteristics in the SEER-Medicare database: Applications and limitations. *Medical Care*, *40*(8), 19-25.

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.

Berry, D., & O'Connor, E. (2010). Behavioral risk, teacher–child relationships, and social skill development across middle childhood: A child-by-environment analysis of change. *Journal of Applied Developmental Psychology*, *31*(1), 1-14.

Beuker, K. T., Schjølberg, S., Lie, K. K., Donders, R., Lappenschaar, M., Swinkels, S. H., & Buitelaar, J. K. (2013). The structure of autism spectrum disorder symptoms in the general population at 18 months. *Journal of Autism and Developmental Disorders*, *43*(1), 45-56.

Billstedt, E., Gillberg, I., & Gillberg, C. (2007). Autism in adults: symptom patterns and early childhood predictors. Use of the DISCO in a community sample followed from childhood. *Journal of Child Psychology and Psychiatry*, *48*(11), 1102-1110.

Bishop, S. L., Guthrie, W., Coffing, M., & Lord, C. (2011). Convergent validity of the Mullen Scales of Early Learning and the differential ability scales in children with autism spectrum disorders. *American Journal on Intellectual and Developmental Disabilities*, *116*(5), 331-343.

Boyd, B. A., Odom, S. L., Humphreys, B. P., & Sam, A. M. (2010). Infants and toddlers with autism spectrum disorder: Early identification and early intervention. *Journal of Early Intervention*, *32*(2), 75-98.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research, Second edition*. Guilford Publications: New York, NY.

Brown, G. L., Craig, A. B., & Halberstadt, A. G. (2015). Parent gender differences in emotion socialization behaviors vary by ethnicity and child gender. *Parenting*, *15*(3), 135-157.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. *Sage Focus Editions*, *21*(2), 136-154.

Bryson, S. E., & Smith, I. M. (1998). Epidemiology of autism: Prevalence, associated characteristics, and implications for research and service delivery. *Mental Retardation and Developmental Disabilities Research Reviews*, *4*(2), 97-103.

Burghardt, G. M. (2011). Defining and recognizing play. In A. Pellegrini (Ed.), *Oxford Handbook of the Development of Play* (pp. 9 –18). New York, NY: Oxford University Press.

Bussey, K., & Bandura, A. (1999). Social cognitive theory of gender development and differentiation. *Psychological Review*, *106*(4), 676-713.

Campbell, S. B., Mahoney, A. S., Northrup, J., Moore, E. L., Leezenbaum, N. B., & Brownell, C. A. (2018). Developmental changes in pretend play from 22-to 34-months in younger siblings of children with autism spectrum disorder. *Journal of Abnormal Child Psychology*, *46*(3), 639-654.

Carlson, S. M., & Taylor, M. (2005). Imaginary companions and impersonated characters: Sex differences in children's fantasy play. *Merrill-Palmer Quarterly*, *51*(1), 93-118.

Carter, A. S., Briggs-Gowan, M. J., Jones, S. M., & Little, T. D. (2003). The infant–toddler social and emotional assessment (ITSEA): Factor structure, reliability, and validity. *Journal of Abnormal Child Psychology*, *31*(5), 495-514.

Cauvet, E., van't Westeinde, A., Toro, R., Kuja-Halkola, R., Neufeld, J., Mevel, K., & Bölte, S. (2018). Sex differences along the Autism Continuum: A twin study of brain structure. *Cerebral Cortex*, *29*(3), 1342-1350.

Chaplin, T. M., & Aldao, A. (2013). Gender differences in emotion expression in children: A meta-analytic review. *Psychological Bulletin*, *139*(4), 735-793.

Charman, T., Swettenham, J., Baron-Cohen, S., Cox, A., Baird, G., & Drew, A. (1997). Infants with autism: An investigation of empathy, pretend play, joint attention, and imitation. *Developmental Psychology*, *33*(5), 781-789.

Cherney, I. D., Kelly-Vance, L., Gill Glover, K, Ruane, A., & Oliver Ryalls, B. (2003). The effects of stereotyped toys and gender on play assessment in children aged 18-47 months. *Educational Psychology*, *23*(1), 95-106.

Christensen, D. L., Bilder, D. A., Zahorodny, W., Pettygrove, S., Durkin, M. S., Fitzgerald, R. T., Rice, C., Kurzius-Spencer, M. Baio, J., & Yeargin-Allsopp, M. (2016). Prevalence and characteristics of autism spectrum disorder among 4-year-old children in the autism and developmental disabilities monitoring network. *Journal of Developmental & Behavioral Pediatrics*, *37*(1), 1-8.

Cole, P. M., Zahn-Waxler, C., & Smith, K. D. (1994). Expressive control during a disappointment: Variations related to preschoolers' behavior problems. *Developmental Psychology*, *30*(6), 835–846

Conlon, O., Volden, J., Smith, I. M., Duku, E., Zwaigenbaum, L., Waddell, C., Szatmari, P., Mirenda, P., Vaillancourt, T., Bennett, T., Georgiades, S., Elsabbagh, M., Ungar, W. J., & The Pathways in ASD Study Team (2019). Gender Differences in Pragmatic Communication in School-Aged Children with Autism Spectrum Disorder (ASD). *Journal of Autism and Developmental Disorders*, *49*(5), 1937-1948.

Connellan, J., Baron-Cohen, S., Wheelwright, S., Batki, A., & Ahluwalia, J. (2000). Sex differences in human neonatal social perception. *Infant Behavior and Development*, *23*(1), 113-118.

Constantino, J. N., Kennon-McGill, S., Weichselbaum, C., Marrus, N., Haider, A., Glowinski, A. L., Gillespie, S., Klaiman, C., Klin, A. & Jones, W. (2017). Infant viewing of social scenes is under genetic control and is atypical in autism. *Nature*, *547*(7663), 340-344.

Dawson, G., Rogers, S., Munson, J., Smith, M., Winter, J., Greenson, J., Donaldson, A., & Varley, J. (2010). Randomized, controlled trial of an intervention for toddlers with autism: The Early Start Denver Model. *Pediatrics*, *125*(1), 17-23.

de Bildt, A., Sytema, S., Zander, E., Bölte, S., Sturm, H., Yirmiya, N., Yaari, M., Charman, T., Salomone, E., LeCouteur, A., Green, J., Bedia R. C., Primo, P. G., van Daalen, E., de Jonge, M. V., Guomundsdottir, E., Johannsdottir, S., Raleva, M., Boskova, M.…Oosterling, I. J. (2015). Autism Diagnostic Interview-Revised (ADI-R) algorithms for toddlers and young preschoolers: Application in a non-US sample of 1,104 children. *Journal of Autism and Developmental Disorders*, *45*(7), 2076-2091.

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.

Delaherche, E., Chetouani, M., Mahdhaoui, A., Saint-Georges, C., Viaux, S., & Cohen, D. (2012). Interpersonal synchrony: A survey of evaluation methods across disciplines. *IEEE Transactions on Affective Computing*, *3*(3), 349-365.

Denham, S. A., Mitchell-Copeland, J., Strandberg, K., Auerbach, S., & Blair, K. (1997). Parental contributions to preschoolers' emotional competence: Direct and indirect effects. *Motivation and Emotion*, *21*(1), 65-86.

DiLavore, P. C., Lord, C., & Rutter, M. (1995). The pre-linguistic autism diagnostic observation schedule. *Journal of Autism and Developmental Disorders*, *25*(4), 355-379.

DiPietro, J. A. (1981). Rough and tumble play: A function of gender. *Developmental Psychology*, *17*(1), 50-58.

Dumont, R., Cruse, C. L., Alfonso, V., & Levine, C. (2000). Book review: Mullen scales of early learning: AGS edition. *Journal of Psychoeducational Assessment*, *18*(4), 381-389.

Dunn, J. (1994). Understanding others and the social world: Current issues in developmental research and their relation to preschool experiences and practice. *Journal of Applied Developmental Psychology*, *15*(4), 571-583.

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.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8*(4), 341-349.

Esler, A. N., Bal, V. H., Guthrie, W., Wetherby, A., Weismer, S. E., & Lord, C. (2015). The autism diagnostic observation schedule, toddler module: standardized severity scores. *Journal of Autism and Developmental Disorders*, *45*(9), 2704-2720.

Fabes, R. A., Martin, C. L., & Hanish, L. D. (2003). Young children's play qualities in same-, other-, and mixed-sex peer groups. *Child Development*, *74*(3), 921-932.

Falkmer, T., Anderson, K., Falkmer, M., & Horlin, C. (2013). Diagnostic procedures in autism spectrum disorders: a systematic literature review. *European Child & Adolescent Psychiatry, 22*(6), 329-340.

Farver, J. A. M., & Wimbarti, S. (1995). Paternal participation in toddlers' pretend play. *Social Development*, *4*(1), 17-31.

Fernell, E., & Gillberg, C. (2010). Autism spectrum disorder diagnoses in Stockholm preschoolers. *Research in Developmental Disabilities*, *31*(3), 680-685.

Fivush, R. (1989). Exploring sex differences in the emotional content of mother-child conversations about the past. *Sex Roles*, *20*(11-12), 675-691.

Fivush, R., Brotman, M. A., Buckner, J. P., & Goodman, S. H. (2000). Gender differences in parent–child emotion narratives. *Sex Roles*, *42*(3-4), 233-253.

Fombonne, E. (2003). Epidemiological surveys of autism and other pervasive developmental disorders: An update. *Journal of Autism and Developmental Disorders*, *33*(4), 365-382.

Forgeot d'Arc, B., Delorme, R., Zalla, T., Lefebvre, A., Amsellem, F., Moukawane, S., Letellier, L., Leboyer, M., Mouren, M., & Ramus, F. (2017). Gaze direction detection in autism spectrum disorder. *Autism*, *21*(1), 100-107.

Freeman, N. K. (2007). Preschoolers' perceptions of gender appropriate toys and their parents' beliefs about genderized behaviors: Miscommunication, mixed messages, or hidden truths? *Early Childhood Education Journal, 34*(5), 357-366.

French, L., & Kennedy, E. M. (2018). Annual Research Review: Early intervention for infants and young children with, or at-risk of, autism spectrum disorder: a systematic review. *Journal of Child Psychology and Psychiatry*, *59*(4), 444-456.

Garside, R. B., & Klimes-Dougan, B. (2002). Socialization of discrete negative emotions: Gender differences and links with psychological distress. *Sex Roles*, *47*(3-4), 115-128.

Giarelli, E., Wiggins, L. D., Rice, C. E., Levy, S. E., Kirby, R. S., Pinto-Martin, J., & Mandell, D. (2010). Sex differences in the evaluation and diagnosis of autism spectrum disorders among children. *Disability and Health Journal*, *3*(2), 107-116.

Goddard, L., Dritschel, B., & Howlin, P. (2014). A preliminary study of gender differences in autobiographical memory in children with an autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *44*(9), 2087-2095.

Goodwin, M. H. (2006). *The Hidden Life of Girls: Games of Stance, Status, and Exclusion*. Malden, MA: Blackwell.

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.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549-576.

Gunnar, M. R., & Donahue, M. (1980). Sex differences in social responsiveness between six months and twelve months. *Child Development, 51*(1), 262-265.

Guralnick, M. J., & Weinhouse, E. (1984). Peer-related social interactions of developmentally delayed young children: Development and characteristics. *Developmental Psychology*, *20*(5), 815-827.

Guthrie, W., Swineford, L. B., Wetherby, A. M., & Lord, C. (2013). Comparison of DSM-IV and DSM-5 factor structure models for toddlers with autism spectrum disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, *52*(8), 797-805.

Hall, J. K., Hutton, S. B., & Morgan, M. J. (2010). Sex differences in scanning faces: Does attention to the eyes explain female superiority in facial expression recognition? *Cognition & Emotion*, *24*(4), 629-637.

Harrison, A. J., Long, K. A., Tommet, D. C., & Jones, R. N. (2017). Examining the role of race, ethnicity, and gender on social and behavioral ratings within the Autism Diagnostic Observation Schedule. *Journal of Autism and Developmental Disorders*, *47*(9), 2770-2782.

Harrop, C., Green, J., Hudry, K., & PACT Consortium. (2017). Play complexity and toy engagement in preschoolers with autism spectrum disorder: Do girls and boys differ? *Autism*, *21*(1), 37-50.

Harrop, C., Jones, D., Zheng, S., Nowell, S. W., Boyd, B. A., & Sasson, N. (2018). Sex differences in social attention in autism spectrum disorder. *Autism Research*, *1*(9), 1264-1275.

Harrop, C., Shire, S., Gulsrud, A., Chang, Y. C., Ishijima, E., Lawton, K., & Kasari, C. (2015). Does gender influence core deficits in ASD? An investigation into social-communication and play of girls and boys with ASD. *Journal of Autism and Developmental Disorders*, *45*(3), 766-777.

Hartung, C. M., & Widiger, T. A. (1998). Gender differences in the diagnosis of mental disorders: Conclusions and controversies of the DSM–IV. *Psychological Bulletin*, *123*(3), 260-278.

Hassett, J. M., Siebert, E. R., & Wallen, K. (2008). Sex differences in rhesus monkey toy preferences parallel those of children. *Hormones and Behavior*, *54*(3), 359-364.

Hedley, D., Nevill, R., Uljarević, M., Butter, E., & Mulick, J. A. (2016). ADOS-2 Toddler and Module 1 standardized severity scores as used by community practitioners. *Research in Autism Spectrum Disorders*, *32*, 84-95.

Herskind, A., Greisen, G., & Nielsen, J. B. (2014). Early identification and intervention in cerebral palsy. *Developmental Medicine & Child Neurology*, *57*(1), 29-36.

Hicks, S. D., Rajan, A. T., Wagner, K. E., Barns, S., Carpenter, R. L., & Middleton, F. A. (2018). Validation of a salivary RNA test for childhood autism spectrum disorder. *Frontiers in Genetics*, *9*(534), 1-11.

Hiller, R. M., Young, R. L., & Weber, N. (2014). Sex differences in autism spectrum disorder based on DSM-5 criteria: evidence from clinician and teacher reporting. *Journal of Abnormal Child Psychology*, *42*(8), 1381-1393.

Hittelman, J. H., & Dickes, R. (1979). Sex differences in neonatal eye contact time. *Merrill-Palmer Quarterly of Behavior and Development*, *25*(3), 171-184.

Howes, C. (1980). Peer play scale as an index of complexity of peer interaction. *Developmental Psychology*, *16*(4), 371-371.

Howlin, P., Goode, S., Hutton, J., & Rutter, M. (2004). Adult outcome for children with autism. *Journal of Child Psychology and Psychiatry, 45*(2), 212-229.

Howlin, P., & Magiati, I. (2017). Autism spectrum disorder: Outcomes in adulthood. *Current Opinion in Psychiatry*, *30*(2), 69-76.

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

Jacquemont, S., Coe, B. P., Hersch, M., Duyzend, M. H., Krumm, N., Bergmann, S., Beckmann, J. S., Rosenfeld, J. A., & 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.

Jadva, V., Hines, M., & Golombok, S. (2010). Infants’ preferences for toys, colors, and shapes: Sex differences and similarities. *Archives of Sexual Behavior*, *39*(6), 1261-1273.

Johnson, K. L., Lurye, L. E., & Tassinary, L. G. (2010). Sex categorization among preschool children: Increasing utilization of sexually dimorphic cues. *Child Development*, *81*(5), 1346-1355.

Killen, M., & Smetana, J. G. (1999). Social interactions in preschool classrooms and the development of young children's conceptions of the personal. *Child Development*, *70*(2), 486-501.

Killen, M., & Turiel, E. (1991). Conflict resolution in preschool social interactions. *Early Education and Development*, *2*(3), 240-255.

Kim, S. H., & Lord, C. (2012). New autism diagnostic interview-revised algorithms for toddlers and young preschoolers from 12 to 47 months of age. *Journal of Autism and Developmental Disorders, 42*(1), 82-93.

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.

Kleberg, J. L., Nyström, P., Bölte, S., & Falck-Ytter, T. (2018). Sex Differences in Social Attention in Infants at Risk for Autism. *Journal of Autism and Developmental Disorders, 49*(4), 1-10.

Knickmeyer, R. C., Wheelwright, S., & Baron-Cohen, S. B. (2008). Sex-typical play: masculinization/defeminization in girls with an autism spectrum condition. *Journal of Autism and Developmental Disorders, 38*(6), 1028-1035.

Kopp, S., & Gillberg, C. (2011). The Autism Spectrum Screening Questionnaire (ASSQ)-Revised Extended Version (ASSQ-REV): an instrument for better capturing the autism phenotype in girls? A preliminary study involving 191 clinical cases and community controls. *Research in Developmental Disabilities, 32*(6), 2875-2888.

Kothari, R., Skuse, D., Wakefield, J., & Micali, N. (2013). Gender differences in the relationship between social communication and emotion recognition. *Journal of the American Academy of Child & Adolescent Psychiatry, 52*(11), 1148-1157.

Krasnor, L. R., & Pepler, D. J. (1980). The study of children's play: Some suggested future directions. In K. H. Rubin (Ed.), Children's Play: New directions for child development (pp. 85–95). San Francisco, CA: Jossey-Bass

Kreiser, N. L., & White, S. W. (2014). ASD in females: are we overstating the gender difference in diagnosis? *Clinical Child and Family Psychology Review*, *17*(1), 67-84.

Li, R. Y. H., & Wong, W. I. (2016). Gender-typed play and social abilities in boys and girls: Are they related? *Sex Roles*, *74*(9-10), 399-410.

Lillard, A. S. (1993). Pretend play skills and the child's theory of mind. *Child Development*, *64*(2), 348-371.

Lindsey, E. W., & Mize, J. (2001). Contextual differences in parent–child play: Implications for children's gender role development. *Sex Roles*, *44*(3-4), 155-176.

Little, T. D., Jorgensen, T. D., Lang, K. M., & Moore, E. W. G. (2014). On the joys of missing data. *Journal of Pediatric Psychology*, *39*(2), 151-162.

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., DiLavore, P. C., & Gotham, K. (2012). *Autism Diagnostic Observation Schedule*. Western Psychological Services.

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., 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. *Journal of Autism and Developmental Disorders*, *30*(3), 205-223.

Lord, C., Rutter, M. L., 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*(2), 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.

Lovaas, O. I. (1987). Behavioral treatment and normal educational and intellectual functioning in young autistic children. *Journal of Consulting and Clinical Psychology*, *55*(1), 3-9.

Luyster, R., Gotham, K., Guthrie, W., Coffing, M., Petrak, R., Pierce, K., Bishop, S., Esler, A., Hus, V., Oti, R., Richler, J., Risi, S., & Lord, C. (2009). The Autism Diagnostic Observation Schedule—Toddler Module: A new module of a standardized diagnostic measure for autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *39*(9), 1305-1320.

Malatesta, C. Z., & Haviland, J. M. (1982). Learning display rules: The socialization of emotion expression in infancy. *Child Development*, *53*(4), 991-1003.

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.

Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, *11*(3), 320-341.

Martin, J. L., & Ross, H. S. (2005). Sibling aggression: Sex differences and parents' reactions. *International Journal of Behavioral Development*, *29*(2), 129-138.

Matson, J. L., & Smith, K. R. (2008). Current status of intensive behavioral interventions for young children with autism and PDD-NOS. *Research in Autism Spectrum Disorders*, *2*(1), 60-74.

Matsuyoshi, D., Kuraguchi, K., Tanaka, Y., Uchida, S., Ashida, H., & Watanabe, K. (2014). Individual differences in autistic traits predict the perception of direct gaze for males, but not for females. *Molecular Autism*, *5*(1), 12-15.

Mattila, M. L., Kielinen, M., Linna, S. L., Jussila, K., Ebeling, H., Bloigu, R., Joseph, R. M., & 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.

McClure, E. B. (2000). A meta-analytic review of sex differences in facial expression processing and their development in infants, children, and adolescents. *Psychological Bulletin*, *126*(3), 424-453.

McCormick, C. E., Kavanaugh, B. C., Sipsock, D., Righi, G., Oberman, L. M., Moreno De Luca, D., Uzun, E., D., Best, C. R., Jerskey, B. A., Quinn, J. G., Jewel, S. B., Wu, P., McLean, R. L., Levine, T P., Tokadijan, H., Perkins, K. A., Lcarke, E. B., Dunn, B., Gerber, A. H., …Morrow, E. M. (2020). Autism Heterogeneity in a Densely Sampled US Population: Results From the First 1,000 Participants in the RI-CART Study. *Autism Research*.

McLennan, J. D., Lord, C., & Schopler, E. (1993). Sex differences in higher functioning people with autism. *Journal of Autism and Developmental Disorders, 23*(2), 217-227.

Milner, V., McIntosh, H., Colvert, E., & Happé, F. (2019). A qualitative exploration of the female experience of autism spectrum disorder (ASD). *Journal of Autism and Developmental Disorders*, *49*(6), 2389-2402.

Mitchell, S., Brian, J., Zwaigenbaum, L., Roberts, W., Szatmari, P., Smith, I., & Bryson, S. (2006). Early language and communication development of infants later diagnosed with autism spectrum disorder. *Journal of Developmental & Behavioral Pediatrics*, *27*(2), S69-S78.

Mullen, E. M. (1995). *Mullen Scales of Early Learning* (pp. 58-64). Circle Pines, MN: AGS.

Mundy, P., Block, J., Delgado, C., Pomares, Y., Van Hecke, A. V., & Parlade, M. V. (2007). Individual differences and the development of joint attention in infancy. *Child Development*, *78*(3), 938-954.

Muthén, L. K., & Muthén, B. O. (2005). *Mplus: Statistical Analysis with Latent Variables: User's Guide* (pp. 1998-2012). Los Angeles, CA: Muthén & Muthén.

Nguyen, T. A., Wu, K., Pandey, S., Lehr, A. W., Li, Y., Bemben, M. A., Badger II, J. D., Lauzon, J. L., Wang, T., Zaghloul, K. A., Thurm, A., Jain, M., Lu, W., Roche, K. W. (2020). A Cluster of Autism-Associated Variants on X-Linked NLGN4X Functionally Resemble NLGN4Y. *Neuron*. In Press https://doi.org/10.1016/j.neuron.2020.03.008

Øien, R. A., Hart, L., Schjølberg, S., Wall, C. A., Kim, E. S., Nordahl-Hansen, A., Eisemann, M. R., Chawarska, K., Volkmar, F. R., & Shic, F. (2017). Parent-endorsed sex differences in toddlers with and without ASD: utilizing the M-CHAT. *Journal of Autism and Developmental Disorders*, *47*(1), 126-134.

Olson, H. C., & Montague, R. A. (2011). An innovative look at early intervention for children affected by prenatal alcohol exposure. In S. A. Adubato & D. E. Cohen (Eds.), *Prenatal Alcohol Use and Fetal Alcohol Spectrum Disorders: Diagnosis, Assessment and New*

Directions in Research and Multimodal Treatment (64-107). Oak Park, IL: Bentham Books.

Osofsky, J. D., & O'Connell, E. J. (1977). Patterning of newborn behavior in an urban population. *Child Development*, *48*(2), 532-536.

Petrou, A. M., Parr, J. R., & McConachie, H. (2018). Gender differences in parent-reported age at diagnosis of children with autism spectrum disorder. *Research in Autism Spectrum Disorders, 50*(2018), 32-42.

Pettit, G. S., & Arsiwalla, D. D. (2008). Commentary on special section on "bidirectional parent–child relationships": The continuing evolution of dynamic, transactional models of parenting and youth behavior problems. *Journal of Abnormal Child Psychology, 36*(5), 711-718.

Pierce, K., Gazestani, V. H., Bacon, E., Barnes, C. C., Cha, D., Nalabolu, S., Lopez, L., Moore, A., Pence-Stophaeros, S., & Courchesne, E. (2019). Evaluation of the Diagnostic Stability of the Early Autism Spectrum Disorder Phenotype in the General Population Starting at 12 Months. *Journal of the American Medical Association: Pediatrics, 173*(6), 578-587.

Pomerleau, A., Bolduc, D., Malcuit, G., & Cossette, L. (1990). Pink or blue: Environmental gender stereotypes in the first two years of life. *Sex Roles, 22*(5-6), 359-367.

Ramsey, R. K., Nichols, L., Ludwig, N. N., Fein, D., Adamson, L. B., & Robins, D. L. (2018). Brief Report: Sex Differences in Parental Concerns for Toddlers with Autism Risk. *Journal of Autism and Developmental Disorders, 48*(12), 4063-4069.

Reichow, B., Barton, E. E., Boyd, B. A., & Hume, K. (2012). Early intensive behavioral intervention (EIBI) for young children with autism spectrum disorders (ASD). *Cochrane Database Systematic Review*, *5*, 1-63.

Rheingold, H. L., & Cook, K. V. (1975). The contents of boys' and girls' rooms as an index of parents' behavior. *Child Development*, *46*(2), 459-463.

Richler, J., Bishop, S. L., Kleinke, J. R., & Lord, C. (2007). Restricted and repetitive behaviors in young children with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *37*(1), 73-85.

Rivet, T. T., & Matson, J. L. (2011). Review of gender differences in core symptomatology in autism spectrum disorders. *Research in Autism Spectrum Disorders*, *5*(3), 957-976.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592.

Sacrey, L. A. R., Zwaigenbaum, L., Szatmari, P., Bryson, S., Georgiades, S., Brian, J., ... & Elsabbagh, M. (2017). Brief report: characteristics of preschool children with ASD vary by ascertainment. *Journal of Autism and Developmental Disorders*, *47*(5), 1542-1550.

Servin, A., Bohlin, G., & Berlin, L. (1999). Sex differences in 1-, 3-, and 5-year-olds' toy-choice in a structured play-session. *Scandinavian Journal of Psychology*, *40*(1), 43-48.

Shattuck, P. T., Durkin, M., Maenner, M., Newschaffer, C., Mandell, D. S., Wiggins, L., Lee, L., Rice, C., Giarelli, E., Kirby, R., Baio, J., Pinto-Martin, J., & 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 & Adolescent Psychiatry*, *48*(5), 474-483.

Shaw, K. A., Maenner, M. J., Baio, J., Washington, A., Christensen, D. L., Wiggins, L. D., Pettygrove, S., Andrews, J. G., White, T., Rosenberg, C. R. Constantino, J. N., Fitzgerald, R. T., Zahorodny, W., Shenouda, J., Daniels, J. L., Salinas, A., Durkin, M. S.

& Dietz, P. M. (2020). Early Identification of Autism Spectrum Disorder Among Children Aged 4 Years—Early Autism and Developmental Disabilities Monitoring Network, Six Sites, United States, 2016. *Morbidity and Mortality Weekly Report. Surveillance Summaries*, *69*(3), 1-11.

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

Sipes, M., Matson, J. L., Worley, J. A., & Kozlowski, A. M. (2011). Gender differences in symptoms of autism spectrum disorders in toddlers. *Research in Autism Spectrum Disorders*, *5*(4), 1465-1470.

Skuse, D. H. (2000). Imprinting, the X-chromosome, and the male brain: explaining sex differences in the liability to autism. *Pediatric Research*, *47*(1), 9.

Skuse, D. H. (2007). Rethinking the nature of genetic vulnerability to autistic spectrum disorders. *TRENDS in Genetics*, *23*(8), 387-395.

Smith, K. A., Gehricke, J. G., Iadarola, S., Wolfe, A., & Kuhlthau, K. A. (2020). Disparities in service use among children with autism: a systematic review. *Pediatrics*, *145*(Supplement 1), S35-S46.

South, M., Ozonoff, S., & McMahon, W. M. (2005). Repetitive behavior profiles in Asperger syndrome and high-functioning autism. *Journal of Autism and Developmental Disorders*, *35*(2), 145-158.

Stanley, G. C., & Konstantareas, M. M. (2007). Symbolic play in children with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *37*(7), 1215-1223.

Thompson, S. K., & Bentler, P. M. (1971). The priority of cues in sex discrimination by children and adults. *Developmental Psychology*, *5*(2), 181-185.

Tierney, S., Burns, J. and Kilbey, E. (2016). Looking behind the mask: social coping strategies of girls on the autistic spectrum. *Research in Autism Spectrum Disorders,* *23*, 73-83.

Tronick, E. Z., & Cohn, J. F. (1989). Infant-mother face-to-face interaction: Age and gender differences in coordination and the occurrence of miscoordination. *Child Development*, *60*(1), 85-92.

Tsai, L., Stewart, M. A., & August, G. (1981). Implication of sex differences in the familial transmission of infantile autism. *Journal of Autism and Developmental Disorders*, *11*(2), 165-173.

Turner, T. N., Wilfert, A. B., Bakken, T. E., Bernier, R. A., Pepper, M. R., Zhang, Z., Torene, R. I., Retterer, K., Eichler, E. E. (2019). Sex-Based Analysis of De Novo Variants in Neurodevelopmental Disorders. *American Journal of Human Genetics*, *105*(6), 1274-1285.

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.

Vismara, L. A., & Lyons, G. L. (2007). Using perseverative interests to elicit joint attention behaviors in young children with autism: Theoretical and clinical implications for understanding motivation. *Journal of Positive Behavior Interventions*, *9*(4), 214-228.

Wang, S., Deng, H., You, C., Chen, K., Li, J., Tang, C., Ceng, C., Zou, Y., & Zou, X. (2017). Sex differences in diagnosis and clinical phenotypes of Chinese children with autism spectrum disorder. *Neuroscience Bulletin*, *33*(2), 153-160.

Weinberg, M. K., Tronick, E. Z., Cohn, J. F., & Olson, K. L. (1999). Gender differences in emotional expressivity and self-regulation during early infancy. *Developmental Psychology*, *35*(1), 175-188.

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, *72*(3), 655-684.

Werling, D. M., & Geschwind, D. H. (2013). Sex differences in autism spectrum disorders. *Current Opinion in Neurology*, *26*(2), 146-153.

Whiteley, P., Todd, L., Carr, K., & Shattock, P. (2010). Gender ratios in autism, Asperger syndrome and autism spectrum disorder. *Autism Insights*, *2010*(2), 17-24.

Whyte, E. M., & Scherf, K. S. (2018). Gaze following is related to the broader autism phenotype in a sex-specific way: building the case for distinct male and female autism phenotypes. *Clinical Psychological Science*, *6*(2), 280-287.

Widaman, K. F. (2006). Best practices in quantitative methods for developmentalists: III. Missing data: What to do with or without them. *Monographs of the Society for Research in Child Development*, *71*(3), 42-64.

Wierenga, L. M., Sexton, J. A., Laake, P., Giedd, J. N., Tamnes, C. K., & Pediatric Imaging, Neurocognition, and Genetics Study. (2017). A key characteristic of sex differences in the developing brain: greater variability in brain structure of boys than girls. *Cerebral Cortex*, *28*(8), 2741-2751.

Wing, L. (1981). Sex ratios in early childhood autism and related conditions. *Psychiatry Research, 5*(2), 129-137.

Zablotsky, B., Colpe, L. J., Pringle, B. A., Kogan, M. D., Rice, C., & Blumberg, S. J. (2017). Age of parental concern, diagnosis, and service initiation among children with autism spectrum disorder. *American Journal on Intellectual and Developmental Disabilities, 122*(1), 49-61.

Zosuls, K. M., Ruble, D. N., Tamis-LeMonda, C. S., Shrout, P. E., Bornstein, M. H., & Greulich, F. K. (2009). The acquisition of gender labels in infancy: Implications for gender-typed play. *Developmental Psychology, 45*(3), 688-717.

Zwaigenbaum, L., Bauman, M. L., Choueiri, R., Kasari, C., Carter, A., Granpeesheh, D., Mailloux, A., Roley, S. S., Wagner, S., Fein, D., Pierce, K., Buie, T., Davis, P. A., Newschaffer, C., Robins, D., Wetherby, A., Stone, W. L., Yirmiya, N., Estes, A. …Natowicz, M. R. (2015). Early intervention for children with autism spectrum disorder under 3 years of age: Recommendations for practice and research. *Pediatrics, 136*(Supplement 1), S60-S81.