J Autism Dev Disord (2017) 47:2770–2782
DOI 10.1007/s10803-017-3176-3

ORIGINAL PAPER

Examining the Role of Race, Ethnicity, and Gender on Social and Behavioral Ratings Within the Autism Diagnostic Observation Schedule

Ashley J. Harrison[1] · Kristin A. Long[2] · Douglas C. Tommet[3] · Richard N. Jones[3]

Published online: 8 June 2017

Abstract The Autism Diagnostic Observation Schedule (ADOS) is widely used to assess symptoms of autism spectrum disorder (ASD). Given well-documented differences in social behaviors across cultures, this study examined whether item-level biases exist in ADOS scores across sociodemographic groups (race, ethnicity, and gender). We examined a subset of ten ADOS items among participants (N = 2458). Holding level of overall ADOS behavioral symptoms constant, we found significant item level bias (measurement noninvariance) for race and ethnicity on three ADOS items. Item-level bias was not apparent across gender. Although the magnitude of bias was small, our findings highlight the need to reevaluate norms and operational definitions used in assessments to increase ASD diagnostic accuracy among culturally-diverse groups.

Keywords Autism spectrum disorder · Cross-cultural · Social norms · ADOS · Measurement bias · Race · Ethnicity

Introduction

Recent research has begun to identify patterns of variability across different cultural groups with regard to autism spectrum disorder (ASD) symptom identification thresholds and symptom profiles emerging on a multitude of ASD screening and symptom measures. In terms of symptom identification, children from ethnic minority groups in the United States (U.S.) (Liptak et al. 2008; Mandell et al. 2009) and Europe (El Bouk et al. 2009) are less likely to receive a diagnosis of ASD than White children, and children from U.S. ethnic minority groups receive diagnostic confirmation significantly later than White children (Mandell et al. 2002). These differences could be influenced by biases in minority referrals (Begeer et al. 2009) but could also reflect a cross-cultural difference in ASD symptom recognition. For example, families from India identify impairments in their children approximately 6–10 months later than U.S. parents due to a range of cultural differences (Daley 2004). Additionally, ASD features in girls are identified later than those in boys in some cultures due to variability in gender roles (Al-Salehi and Al-Hifthy 2009). Taken together, there is increasing evidence that variability in ASD rates could relate to the challenges associated with developing psychometrically sound cross-cultural measures for assessing heterogeneous ASD symptoms in children with autism (Soto et al. 2014) as well as neurotypical populations (e.g., Magiati et al. 2015).

Compared to children with ASD from the ethnic majority in the U.S., children from ethnic minority backgrounds have a different symptom profile consisting of lower scores in language and communication abilities and lower cognitive composite scores (Chaidez et al. 2012; Landa and Garrett-Mayer 2006). White children are more likely to have certain ASD symptoms including inflexible adherence to nonfunctional routines/rituals and persistent preoccupation with parts of objects compared to Black children (Sell et al. 2012). With regard to gender, research has shown that caregivers and teachers

✉ Ashley J. Harrison
Ashley.Harrison@uga.edu

[1] Department of Educational Psychology, University of Georgia, Aderhold Hall 110 Carlton Street, Athens, GA 30602, USA

[2] Boston University, Boston, MA, USA

[3] Alpert Medical School of Brown University, Providence, RI, USA

report different initial pre-diagnostic concerns for males and females (Hiller et al. 2015) and that in different cultural contexts males score systematically higher than females on a standardized ASD assessment measure (Freeth et al. 2013). From a more global perspective, cross-country differences have been documented on standardized tests of sensory functioning (Caron et al. 2012), social skills (Sipes et al. 2012), challenging behaviors (Chung et al. 2012) and ASD symptoms (Freeth et al. 2013; Matson et al. 2017, 2011). Without a careful examination of current diagnostic measures, it is not clear if these findings reflect true differences in symptom presentation across sociodemographic groups or if the differences are a function of inherent cross-cultural biases within diagnostic measures.

ASD diagnostic instruments commonly include an evaluation of basic social behaviors to identify the core symptoms of social communication impairment (American Psychiatric Association 2013). Most ASD-specific diagnostic instruments are designed to identify social deficits through a comparison to operationally-defined normative social behavior. For example, the criteria for social-emotional reciprocity is assessed by first observing the facial expressions and shared enjoyment exhibited by the person being evaluated and then determining the extent to which the person's observed social behaviors match the exemplar defined based on norms (e.g., Constantino and Gruber 2012; Lord et al. 2012). Similarly, ASD measures often include items assessing deficits in non-verbal communication through the quantification of more unitary social behaviors such as eye contact or gestures (Mundy 1995). These norm-based comparisons are the cornerstone of gold standard ASD diagnostic assessment (Ozonoff et al. 2005); however, it is exceedingly difficult to account for the wide variability in social norms across cultures. This approach to measuring social functioning may inadvertently lead to the formulation of operational definitions that are based on the majority culture of the U.S. (i.e., Western males) and lack sensitivity to cross-cultural variability, which may introduce the possibility of culture-based biases in assessment instruments (Freeth et al. 2014). In turn, biased assessments may account for some of the observed cross-cultural variability in ASD symptom profiles. Cross-cultural variability may particularly impact how social behaviors are rated when the cultural background of the rater differs from that of the participant. To further illustrate the potential impact of culture and gender on aspects of social behavior assessed by ASD diagnostic instruments, the following sections outline the literature documenting variability across eye contact, facial expression, play, and language.

Cultural Differences in Eye Contact

Cross-cultural variability regarding the amount and type of eye contact during social interactions is well documented (e.g., Fugita et al. 1974; McCarthy et al. 2006) and may, in part, depend on whether the cultural context is classified as individualist versus collectivist (Knapp et al. 2013). McCarthy et al. (2006) observed that Japanese participants demonstrated less eye contact when responding to questions compared to Trinidadian and Canadian participants. These findings align with theories that cross-cultural differences in eye contact stem from variability in how eye contact is valued and interpreted across societies (Knapp et al. 2013). In some cultural contexts (e.g., Kenya), children are discouraged from making eye contact with adults as a sign of respect (Carter et al. 2005), while maintaining constant eye contact during communication is customary in other cultures (Collett 1971). Variability in eye contact might also relate to differences in basic visual processing across cultural contexts. For example, research shows cross-cultural variability in patterns of how participants visually scan social images (Chua et al. 2005).

Differences in eye contact also arise within different demographic groups within the U.S. In an early study examining cross-cultural variability in eye contact, Fugita et al. (1974) found that White students in the U.S. were more likely than Black students to maintain eye contact during an interview regardless of the race of the interviewer. Further, eye contact between White and White dyads is greater than that between Black and Black or Black and White dyads (LaFrance and Mayo 1976). Similarly, European American parents and children demonstrate increased rates of shared eye gaze and heightened overall amounts of child gaze directed at parents compared to members of Mexican American families (Schofield et al. 2008). Given the central role of eye contact in ASD assessment, cross-cultural variability may impact the degree to which current assessment instruments can identify abnormal levels of eye contact if such differences are not reflected in scoring procedures or operational definitions of typical and divergent behavior.

Cultural Differences in Facial Expressions & Recognition

There is also evidence of cross-cultural variability across facial expressiveness and recognition (Elfenbein 2013; Elfenbein et al. 2007). Using computer technology to map mental representations of facial expressions, Jack et al. (2012) found that Easterners are less likely to have a distinct set of facial expressions to represent each of the basic emotions as compared to their Western counterparts. Different cultural groups also exhibit distinct non-verbal facial cues when expressing emotions (Marsh et al. 2003) as well as variability in the method of visually processing of facial cues to interpret facial expressions (Yuki et al. 2007).

Distinctions in facial expressiveness exist even among ethnic groups within the same country. For example, Black participants are generally more emotionally expressive than White participants (Vrana and Rollock 2002). These differences in facial expression may stem from different value systems within these cultural groups. One hypothesis is that Black culture encourages a greater range of expression than White culture because of a heightened value placed on spontaneous expression of feelings and self-assertion (Kochman 1981). In turn, these nuanced multicultural differences may affect more overt behavioral differences in facial expression.

Cultural Differences in Language

Distinct non-verbal cues during communication and different rules used to construct language may impact how a social partner from a different cultural context interprets communication. Anthropologists distinguish high-context cultures that rely more heavily on indirect and implicit messages from low-context cultures that emphasize verbal messages (Knapp et al. 2013). To demonstrate these distinct styles, European and Middle Eastern individuals demonstrate differences in proximity, volume, and non-verbal cues such as touching and body orientation during conversation (Collett 1971). There are also cultural differences in pragmatic language. Discourse rules like turn-taking, interrupting, appropriate topics, and humor are often determined by culture (Carter et al. 2005). Additionally, the frequency of conversations between children and adults varies across cultural contexts. For example, children in rural Kenya may less frequently sit and talk with adults (Carter et al. 2005). This difference may impact how a child performs on instruments rating conversations and interactions with an adult examiner such as the Autism Diagnostic Observation Schedule (ADOS). Finally, for bilingual children, linguistic variability across cultures in terms of pronoun usage may also impact the rating of language performance on ASD assessment scales (Carter et al. 2005).

Cross-cultural differences in language usage may also be associated with different dialects of English spoken in the U.S. Distinct grammatical patterns between different racial groups in the U.S. suggests that variability in language or vocabulary may depend on the degree of exposure to Standard English in one's environment (Hall and Freedle 1975). Regarding proficiency with different English dialects (Standard English compared to Black Vernacular English), children perform better when the testing language matches the language spoken at home (Hall et al. 1975). These dialect differences may have implications for diagnostic testing that utilizes semi-structured interviewing techniques to evaluate conversational back and forth (e.g., ADOS Modules 3 and 4).

Cultural Differences in Play

An examination of play across 20 societies demonstrated variability regarding type of play (i.e., imaginary versus functional) and objects used in play (Ember and Cunnar 2015). Anthropological research by Lancy (2007) also demonstrates variability in play partners across cultures, such that in some cultures there is limited play with parents and extremely rare play with unfamiliar adults. To emphasize these differences in cross-cultural play, Lancy (2007) explains that "psychologists see mother–child play as natural; anthropologists see it as cultural (p. 273)."

There is less research examining play differences among cultural groups within the U.S.; however, Kochman (1981) theorizes that the same cross-cultural differences in self-assertion between Black and White children that may influence facial expression may also impact play styles. For example, more serious, methodical, and purposeful play seems to be more highly valued in White culture, while more unrestrained types of play seems to be viewed more positively in Black culture. Similarly, European-American mothers appear to engage in symbolic and pretend play, while play in Mexican American mother/child dyads resembled more of a shared work activity (Farver and Howes 1993).

Gender Differences in Behavioral Interactions

Socio-cultural gender roles vary across cultural contexts and play an important role in differential social development (Lai et al. 2015). Gender influences social behavioral norms including play styles and emotion expression both among typically developing children and those with ASD. For example, boys tend to play more actively and with construction materials and girls engage in more pretend play routines with dolls and house toys (Knickmeyer et al. 2008). With regard to facial expression, males and females demonstrate differences in overall facial expressiveness (Vrana and Rollock 2002) and in the number of different emotions expressed (Hess et al. 2005; LaFrance and Hecht 2000). A meta-analysis also documents gender differences in facial expression processing attributable to variable early exposure to modeling and parent scaffolding (McClure 2000). The emotional interpretation of facial expression varies depending on gender, such that male faces are usually associated with anger whereas female faces are usually associated with surprise (Zebrowitz et al. 2010).

Potential Impact of Cross-Cultural Variability in ASD Diagnostic Assessment

Taken together, cultural and gender-based variability in social and communication behaviors indicate that the concept of "normal" is not consistent across all cultural contexts whether referring to culture differences in terms of race, ethnicity or gender. It is important to consider how this variability may impact ASD assessment, which is highly reliant on quantifying how observed behaviors deviate from an operational definition of "normal" across social and communication domains. Specifically, operational definitions used in ASD diagnostic measures must account for the wide variability in social norms across cultures to avoid over- or under-identifying social impairment as a result of cultural variability. A review by Kirkovski et al. (2013) discusses how the gender invariance observed in the prevalence of ASD might be in part due to a failure to recognize culturally variant profiles of ASD.

The ADOS is a diagnostic measure that rates observed social behavior as impaired or not based on operational definitions of behavior ranging from typical to consistent with ASD. As evidenced by its translation into 25 languages (Western Psychological Services 2016), the ADOS has wide use across cultural contexts. Although the ADOS includes items that examine social behaviors with cross-cultural variability (e.g., eye contact and facial expression), little research has examined the potential impact of cross-cultural variability in social behaviors on ADOS diagnostic outcomes (Norbury and Sparks 2013). Thus, the goal of this research was to examine potential item-level bias in ADOS items according to race (Caucasian, African American, or Asian), ethnicity (Hispanic or non-Hispanic), and gender (male or female) in the U.S. Our specific question was to determine if there is evidence of differential item functioning (DIF) by group.

Method

Participants

We analyzed data from the Simons Simplex Collection (SSC), which includes phenotypic and genetic information on simplex families (one child affected with ASD) across North America. Participants aged 4–18 were included in the SSC if they were the only child in the family who met criteria for ASD based on scores on standardized ASD diagnostic instruments (Fischbach and Lord 2010). Given inconsistent clinical ratings across sites (Fischbach and Lord 2010), participants were not provided DSM-IV-TR labels (American Psychiatric Association 2000) but were included in the sample if they were identified as having an ASD, as consistent with the proposed changes to DSM-5 (American Psychiatric Association 2013). Participants were recruited from 13 university clinics in the U.S. and Canada specializing in ASD evaluations [Michigan, Yale, Emory, Columbia, Vanderbilt, McGill, Washington, and Harvard Universities and the Universities of Washington, Illinois (Chicago), Missouri, California (Los Angeles), and the Baylor College of Medicine]. All parents in the SSC completed university Institutional Review Board approved informed consents. Additional details regarding inclusion and exclusion criteria can be found in the SFARI Base/SSC Researcher Welcome Packet (Simons 2010) or in a previous study describing the study methodology in detail (Fischbach and Lord 2010).

Participants included in the current analyses (n = 2459) had complete data and fell within one of the racial categories identified as sufficiently powered for analysis: White (n = 2245), Black/African American (n = 103), or Asian (n = 111). Race (White, Black/African, and Asian) and ethnicity (Latino and non-Latino) categories align with those of the U.S. Census Bureau and the National Institutes of Health (Richesson et al. 2014; United States Census Bureau 2017). The majority of participants classified their ethnicity as not Hispanic or Latino (n = 2165) rather than Hispanic or Latino (n = 294) and were male (n = 2129). See Table 1 for participant descriptive data.

Measures

SSC participants completed the original version of the ADOS (Lord et al. 2002), and we examined a subset of ten items from this measure. The ADOS is designed to rate behaviors commonly associated with ASD using a semi-structured, observation-based approach. Each item assessing social communication or repetitive and stereotyped behaviors has a detailed operational definition describing the item theme. Possible scores on items range from either 0–2 or 0–3 (raw scores were not converted to algorithm scores), with higher scores representing a profile more consistent with behaviors observed among individuals with ASD. For each score option, operational definitions are provided to serve as a coding anchor. The ADOS has been widely studied and demonstrates overall excellent psychometric properties (Lord et al. 2000; Mazefsky and Oswald 2006).

To increase our study power, we included participants that had completed any of the four ADOS modules. ADOS items were included in analyses if they were (a) part of at least three of the four ADOS Modules and (b)

Table 1 Participant characteristics (N = 2458)

Characteristic	Overall		African American		Hispanic		Asian		White	
Number of observations	2458	(100)	95	(100)	293	(100)	111	(100)	1959	(100)
Sex [n (%)]										
Female	330	(13.4)	14	(15)	34	(11.6)	14	(12.6)	268	(13.7)
Male	2128	(86.6)	81	(85)	259	(88.4)	97	(87.4)	1691	(86.3)
Age (years) [mean (SD)]	9.06	(3.58)	8.8	(3.2)	8.4	(3.2)	8.5	(3.3)	9.21	(3.66)
Father's education [n (%)]										
Less HS degree	309	(12.7)	12	(13)	63	(22)	4	(3.6)	230	(11.8)
HS degree	721	(29.6)	25	(27)	57	(19.9)	59	(53.6)	580	(29.8)
Some college/associate degree	644	(26.5)	30	(32)	108	(37.6)	15	(13.6)	491	(25.3)
College degree	760	(31.2)	26	(28)	59	(20.6)	32	(29.1)	643	(33.1)
Mother's education [n (%)]										
Less HS degree	205	(8.4)	9	(9)	47	(16.1)	6	(5.5)	143	(7.3)
HS degree	650	(26.5)	27	(28)	60	(20.5)	48	(43.6)	515	(26.4)
Some college/associate degree	713	(29.1)	26	(27)	96	(32.9)	20	(18.2)	571	(29.2)
College degree	883	(36)	33	(35)	89	(30.5)	36	(32.7)	725	(37.1)
Household income [n (%)]										
1: <$20k	68	(2.9)	6	(7)	15	(5.3)	6	(5.7)	41	(2.2)
2: $21–35k	112	(4.8)	8	(9)	21	(7.4)	5	(4.7)	78	(4.2)
3: $36–50k	202	(8.7)	15	(16)	34	(12)	4	(3.8)	149	(8)
4: $51–65k	259	(11.1)	9	(10)	36	(12.7)	9	(8.5)	205	(11.1)
5: $66–80k	330	(14.1)	14	(15)	37	(13.1)	9	(8.5)	270	(14.6)
6: $81–100k	404	(17.3)	11	(12)	50	(17.7)	16	(15.1)	327	(17.6)
7: $101–130k	354	(15.2)	12	(13)	36	(12.7)	20	(18.9)	286	(15.4)
8: $131–160k	217	(9.3)	6	(7)	22	(7.8)	11	(10.4)	178	(9.6)
9: over $161k	389	(16.7)	10	(11)	32	(11.3)	26	(24.5)	321	(17.3)
IQ [mean (SD)]	81.4	(27.7)	64	(24)	76	(25)	72	(32)	83.6	(27.6)
ADOS module [n (%)]										
1	433	(17.6)	31	(33)	71	(24.2)	40	(36.0)	291	(14.9)
2	544	(22.1)	32	(34)	83	(28.3)	32	(28.8)	397	(20.3)
3	1411	(57.4)	32	(34)	136	(46.4)	38	(34.2)	1205	(61.5)
4	70	(2.8)	0	(0)	3	(1.0)	1	(0.9)	66	(3.4)

worded the same across all ADOS Modules in the scoring protocol, both in terms of the general item description and the operational definitions accompanying each score category. For example, items assessing Overall Level of Non-Echoed Spoken Language and Imagination/Creativity had consistent item descriptions but different operational definitions across modules and were therefore omitted. There were two exceptions to retaining questions not syntactically verbatim: (a) items including an additional clause to specify interactions with either the examiner or caregiver and (b) when the content of examples was altered to align with developmental level without altering the root question or response option content. Items were included from both the Social Affect and Restricted and Repetitive Behavior ADOS domains. We excluded items included in the "E" category that measure behaviors associated with disorders commonly comorbid with ASD that may impact performance on the ADOS (i.e., Overactivity; Tantrums, Aggression, Negative or Disruptive Behavior; and Anxiety) but do not specifically assess ASD symptoms.

Based on these selection criteria, ten ADOS items were retained. The majority of these items are included in the revised diagnostic coding algorithms (see Table 2; Gotham et al. 2008). We examined the following items: (1) Stereotyped/Idiosyncratic Use of Words or Phrases, (2) Unusual Eye Contact, (3) Facial Expression Directed to Others, (4) Quality of Social Overtures, (5) Quality of Social Response, (6) Unusual Sensory Interest in Play Material/Person, (7) Hand and Finger and Other Complex Mannerisms, (8) Immediate Echolalia, (9) Overall Quality of Rapport, and (10) Self-Injurious Behavior. See Table 2 for a more detailed description of the examined ADOS items, factor assignment, and module/algorithm inclusion.

Table 2 ADOS items included in analyses

ADOS item	Factor	Module 1	Module 2	Module 3	Module 4
Unusual eye contact	SA	A	A	A	A
Facial expression directed to others	SA	A	A	A	A
Quality of social overtures	SA	A	A	A	A
Quality of social response	SA		X	A	A
Overall quality of rapport	SA		A	A	A
Immediate echolalia	RRB	X	X	X	X
Stereotyped/idiosyncratic use of words or phrases	RRB	A	A	A	A
Unusual sensory interest in play material/person	RRB	A	A	A	A
Hand and finger and other complex mannerisms	RRB	A	A	A	A
Self-injurious behavior	RRB	X	X	X	X

X included in module, *A* included in module and an algorithm item, *RRB* restricted and repetitive behavior, *SA* social affect

Statistical Analysis

We used item response theory (IRT) methods within a structural equation modeling framework to evaluate the psychometric properties of the harmonized ADOS items. IRT describes a family of statistical models developed to describe the internal characteristics of tests and measures. The specific model we used is referred to as a graded response model, with covariates. The model can also be described as a MIMIC (Multiple Indicators, Multiple Causes) model. This description implies that we used multivariate logistic regression for ordered response variables, with latent variables. All parameter estimates were obtained with Mplus software (Version 7.4; Muthén and Muthén 2005). All scripts and detailed output are available upon request.

The analytic approach first confirms a reasonable latent variable model to account for the covariation among selected ADOS items, and then determines if the measurement parameters of that latent variable model differ significantly across gender, race, or ethnicity. The latent variable measurement model is a form of confirmatory factor analysis suitable for categorical dependent variables. The measurement slopes (akin to factor loadings) and item thresholds (boundaries between response categories) are the basic measurement model parameters that are of interest for describing DIF. If these parameters are different across group, it implies that a particular item may be more or less relevant for characterizing increases in underlying severity and/or that the level of severity on the latent trait is differentially represented by endorsement of particular symptoms in a specific group. However, before assuming a group difference in measurement properties exists, it is necessary to rule out if the difference may be attributable to a third variable (i.e., age or cognitive scores), and therefore, we included these as covariates with the confirmatory factor analysis model.

With multiple covariates and multiple outcomes (the ADOS symptoms), we estimated a MIMIC model. We used the multiple group formulation of the MIMIC model to test for differences across gender, race, and ethnicity subgroups. Details on the modeling approach and technical references are described elsewhere (Jones 2006). Briefly, the approach began with a multiple group confirmatory factor analysis with covariates, where the grouping variable was the variable for which we evaluated the presence of DIF. Initially, all measurement parameters (thresholds, loadings, and variances) were held equal across groups. Latent variable intercepts are allowed to vary across groups to allow for group heterogeneity. Item by item, we allowed for thresholds and loadings to vary across groups and collect Chi square difference tests using nested model likelihood ratio tests. We identified model modifications to measurement model parameters with the greatest improvement in model fit, and accept that as evidence of measurement noninvariance and possible DIF if statistically significant ($\alpha = 0.05$). The model with the mean and thresholds free by group so identified became the new null model, and the item by item specification search process was repeated until no statistically significant model modifications were found. The impact of DIF was suggested by comparison of the DIF-naïve (i.e., initial) and DIF-controlled (i.e., final) model group mean differences in the underlying latent traits.

We evaluated the fit of our models using standard methods. Overall model fit was assessed with the root mean square error of approximation (RMSEA; Browne et al. 1993) and the comparative fit index (CFI; Bentler 1990). The RMSEA provides a measure of discrepancy per model degree of freedom. The RMSEA will approach 0 as model fit improves. Hu and Bentler (1998) suggest that values close to 0.06 or less represent adequately fitting models. The CFI is based on the model Chi square, with values that range between 0 and 1. CFI values greater than 0.95 are generally accepted as describing adequately fitting models (Hu and Bentler 1998). The significance of group differences in measurement model was informed with nested model comparisons and computation of a model-Chi square difference test (Satorra 2000). The Mplus robust maximum likelihood estimator was used. This model includes all observations, including those with missing data, and invokes the missing at random (MAR) assumption, which is a less restrictive assumption than what would be possible under a missing completely at random (MCAR) analysis model (e.g., pairwise complete under multivariate probit weighted least squares or analysis of complete data only). This is an important analysis feature, because some module respondents are missing for analysis variables because a representative question was not included in the module. Under our analysis approach, we theoretically obtain unbiased estimates of other item parameters using maximum likelihood.

Results

The response frequencies for the analysis variables are illustrated in Table 3. As can be seen, there was wide variability in the response frequencies across items. In the summary item assessing *Quality of Social Response*, only about 5% of participants were rated in the "0" category, indicating that very few participants showed a range of appropriate social responses. Conversely, about 94% of participants were rated in the "0" category for the *Self Injurious Behavior* item, indicating that it was much more rare for individuals to be rated as exhibiting this symptom during the ADOS.

Results of the confirmatory factor analysis are reported in Table 4. In line with previous research (Gotham et al. 2008), a two factor model provided optimal fit (CFI = 0.932, RMSEA = 0.044), and each was indicated by five items with loadings between 0.4 and 0.83 on their respective

Table 3 Item response frequencies (entries are counts)

Item	Responses					Total
	0	1	2	3	NA	
y1: overall quality of rapport	350	1137	514	24	433	2458
y2: quality of social response	121	1407	480	17	433	2458
y3: quality of social overtures	165	1580	656	57	0	2458
y4: facial expressions directed to others	298	1691	469	0	0	2458
y5: unusual eye contact	210	0	2248	0	0	2458
y6: self-injurious behavior	2311	101	46	0	0	2458
y7: stereotyped/idiosyncratic use of words or phrases	337	1025	828	37	231	2458
y8: unusual sensory interest in play material/person	912	729	817	0	0	2458
y9: hand and finger and other complex mannerisms	1389	404	665	0	0	2458
y10: immediate echolalia	1502	561	283	11	101	2458

NA not available

Table 4 Results of confirmatory factor analysis

Item	Label	Standardized factor loading	Threshold 1	Threshold 2	Threshold 3
F1 by					
y1	Overall quality of rapport	0.79	−0.74	0.74	2.32
y2	Quality of social response	0.83	−1.31	0.84	2.49
y3	Quality of social overtures	0.79	−1.20	0.88	2.39
y4	Facial expressions directed to others	0.55	−1.05	0.99	
y5	Unusual eye contact	0.40	−1.14		
F2 by					
y6	Self-injurious behavior	0.44	1.87	2.41	
y7	Stereotyped/idiosyncratic use of words or phrases	0.54	−0.79	0.44	2.21
y8	Unusual sensory interest in play material/person	0.58	0.01	0.78	
y9	Hand and finger and other complex mannerisms	0.57	0.42	0.89	
y10	Immediate echolalia	0.73	0.67	1.48	2.90

Model confirmatory fit index (CFI) = 0.93, root mean squared error of approximation (RMSEA) = 0.044

factors. The factor correlation was 0.49. Fit of the two factor model was superior to a one factor model (CFI = 0.825, RMSEA = 0.067). Thus, we determined that a two factor model fit the data well and mapped on to ADOS domains as well as DSM 5 symptom clusters of social communication and repetitive and stereotyped behaviors. This allowed us to examine these two factors independently as the latent traits instead of using total ADOS or algorithm scores. Investigating the dimensions independently is particularly important given the heterogeneity observed in ASD symptom profiles.

Results of the DIF detection are summarized in Fig. 1, panels a–d. Four items were found to have significant DIF. The first (panel a, top panel) was *Unusual Eye Contact*, which was different between White and Black participants. The trace function or item characteristic curve (ICC) is plotted for White participants (smooth heavy black line) and Black participants (smooth heavy red line). The observed mean item response is shown with a light line of matching color. The purpose of showing both the model-implied and observed item response functions is to give a visual impression of model fit. We also show with box-and-whisker plots the distribution of underlying trait scores, illustrated in matching colors. Fitted ICCs are only drawn over the range of observed latent trait scores (derived from the fitted model using Bayesian plausible values). The vertical reference line illustrates the location of the item threshold(s). Panel a illustrates that, at a given level of underlying severity on the first latent factor underlying the ADOS, Black participants are rated more highly (i.e., more impaired) on the *Unusual Eye Contact* item.

Panel b of Fig. 1 illustrates the second item found with DIF, which is the same as the first item (*Unusual Eye Contact*), but the group comparison is Hispanic or Latino participants relative to White participants. As with Black participants, Hispanic participants are rated higher on the *Unusual Eye Contact* item than are Whites when comparing at the same level of the underlying trait (Social Communication). Panel c illustrates that Black participants are rated more highly on the *Stereotyped/Idiosyncratic Use of Words or Phrases* item than are White participants when compared at the same level of the second factor (Repetitive and Stereotyped Behaviors), and are rated higher (panel d) on the *Immediate Echolalia* item than comparably severe White respondents.

We assessed the impact of DIF in two ways. First, we examined the estimated difference in latent trait levels across race and ethnicity group with and without statistical control for DIF. In the parlance of the multiple group MIMIC model, we compared the estimated means for the latent trait with and without including the detected threshold and slope differences across race/ethnicity group. This evaluation of DIF impact provided an estimate of the study-wise bias incurred by ignoring measurement differences. The second way we evaluated the impact of DIF was by contrasting individual participant's estimated latent trait levels with and without controlling for DIF. Whereas the study may find minimal impact of DIF on average, the impact for some individual participants may be large. To make this comparison, we contrasted estimated latent trait levels (obtained using *modal a posteriori* factor score estimates) with and without control for DIF.

For the group-level impact analysis, we found that the DIF-naïve difference in factors 1 and 2, respectively, for Black participants was 0.068 and 0.215 without controlling for DIF, and was 0.048 and 0.091 with controlling for DIF. Thus, the impact of controlling for DIF would change our inference that there existed a small to medium effect size difference in the mean level of factor 2 between White and Black participants when we do not control for DIF. However, when we control for DIF, the racial group difference is in the trivial to small range. We judge this to be a non-trivial difference in interpretation. For Hispanic or Latino participants relative to White participants, the DIF-naïve vs DIF-adjusted mean differences were −0.186 (F1) and 0.111 (F2) vs. −0.210 and 0.110, respectively. In this case, we do not observe a substantial impact on the judgment of group differences when DIF is controlled.

For the individual level DIF analysis, none of the Black participants had more than a trivial (<|0.2| standardized units) difference in their estimated level on the first latent trait (social communication), when comparing the DIF-adjusted to DIF-naïve factor scores on the first trait. However, about three quarters of the Black participants had a difference in their estimated level on the second trait (repetitive and stereotyped behaviors) between the conditions where DIF was ignored compared to where DIF was modeled. This individual level impact analysis reinforces the group impact analysis.

Discussion

A measurement noninvariance analysis was used and approached using a MIMIC model analytic framework to examine bias on ADOS items for ethnicity, race, and gender. Holding ADOS subdomain (Social Communication; Repetitive and Stereotyped Behaviors) scores constant, we found significant item level bias according to race and/or ethnicity for three items of the ten ADOS items examined: *Unusual Eye Contact, Stereotyped/Idiosyncratic Use of Words or Phrases*, and *Immediate Echolalia*. More specifically, Black children were more likely to have higher (i.e., more atypical) ratings on the ADOS items assessing levels of *Unusual Eye Contact, Stereotyped or Idiosyncratic Word Use*, and *Immediate Echolalia*.

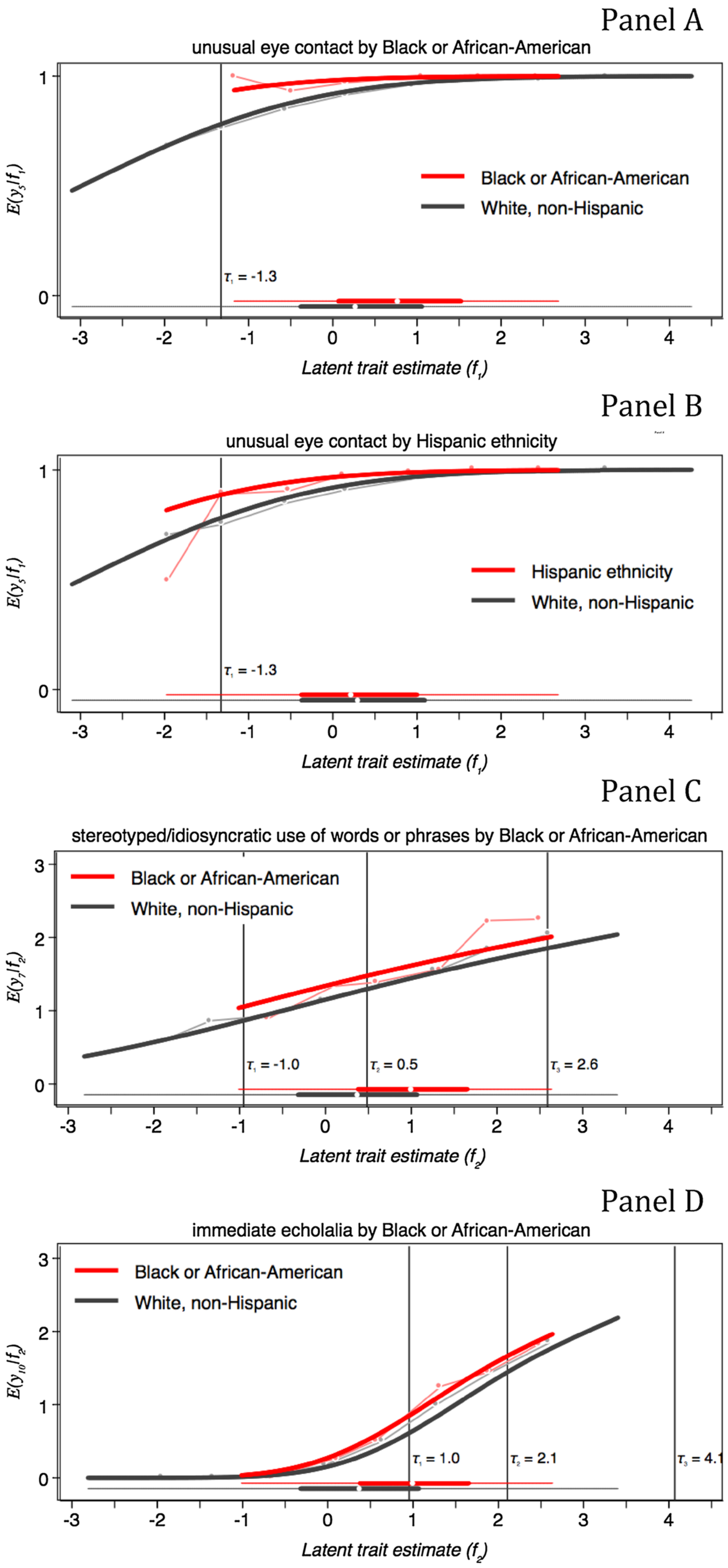

Fig. 1 Results of differential item functioning (DIF) detection. *Each panel* illustrates an expected item response function, which plots the expected (i.e., model-implied) value on the response item (*y-axis*) as a function of the underlying latent trait level (*x-axis*). The reference group is represented by a *heavy black line* and the focal group by a *heavy red line*. We also illustrate the observed mean response (*thin and not smoothed lines*) as an indication of overall model fit. Above the *y-axis*, box-and-whisker plots show the range of estimated latent trait levels for the reference (*black*) and focal groups (*red*). Finally, we illustrate the location of item response level thresholds with *vertical bars*. The extent to which the smoothed response functions separate from each other for the reference and focal group describes the magnitude of the differential item functioning. The top line in each figure aligns with the racial/ethnic minority group. (Color figure online)

In terms of ethnicity, Hispanic children were also more likely to have higher ratings on the item assessing levels of *Usual Eye Contact*. No item level biases were observed for gender. In a diagnostic assessment context, this variability within ADOS items may result in overestimation of impairment for Black and Hispanic groups.

Importantly, the latent trait (ADOS subdomain score) would only be impacted with a systematic difference observed across items. While this wouldn't be the case for Hispanic ethnicity since only one of the examined items demonstrated a bias, identifying multiple items biased for Black children indicates that the problem may be more pronounced for this particular group. In a closer examination to see how the biased items might impact the total score, we investigated whether the biased items appeared in the algorithm. For Modules 1, 2, and 3, only 2 of 14 algorithm items demonstrated a bias; for Module 4, only one of ten algorithm items demonstrated a bias. Therefore, it appears unlikely that in this sample the overall score is being impacted by item-level bias. Although this study found a minimal impact of DIF when averaged across participants, it is possible that this bias may impact diagnostic decisions about individual children (not examined here). This possibility highlights the relevance of the current findings in clinical settings, where clinicians should use caution when applying these three items to children from minority backgrounds.

Further, characteristics of the sample may increase the likelihood that we have underestimated the potential item bias. Data were collected from families who self-referred to a university setting, which has a documented referral bias in favor of White, middle to high SES families as compared to those seen in community settings (Begeer et al. 2009). Thus, although our initial findings suggest that there might be a relatively low impact of item-level bias on ASD assessment on average, these results suggest the need for a more careful examination of how cultural variability in social behaviors may impact diagnostic measures reliant on operational definitions anchored in one cultural context.

Although our overall sample was large, the number of participants identifying as racial or ethnic minorities or female was relatively small. To increase the power of this study, we only examined items that were equivalent across all ADOS Modules; however, this required the omission of multiple items from the diagnostic algorithms. This is a notable limitation of the study, as it prohibited an examination of how the total ADOS score might be impacted by item level-biases. Future investigations with access to larger samples of diverse individuals with ASD may want to examine all items included in the ADOS scoring algorithm to allow for an examination of Differential Test Functioning (DTF; Pae and Park 2006; Runnels 2013) to better understand the impact of cultural variability in social behaviors on ADOS outcomes.

Given the variability in social behaviors between males and females, we were surprised that none of the examined ADOS items, particularly *Facial Expression Directed to Others*, demonstrated a gender bias. It is possible that an item examining a different aspect of facial expression (more an emphasis on quality than quantity) or the inclusion of items assessing play styles may have resulted in different findings with regards to gender biases. Although the sample was insufficiently powered to look at gender-by-race/ethnicity interactions, given the low numbers of girls and racial minorities, this may be an additional area of emphasis in future research. While we were able to examine whether race, gender, and ethnicity may impact ADOS scores, we realize this presents a narrow examination of cultural variables that may impact social interaction. Future investigations would benefit from further unpacking the variables associated with one's culture (e.g., socioeconomic status, birth country, immigration history, language spoken, and acculturation) that may impact how a person interacts in a social context. Similarly, future research may benefit from consideration of the cultural background of the examiner or the cultural match between examiner and examinee.

While the field has made significant gains in providing increased global access to measures like the ADOS through translation (Western Psychological Services 2016), we must emphasize the importance of cultural as well as linguistic adaptation in readying measures for cross-cultural use (Bracken and Barona 1991). As the ADOS has become an important component of thorough ASD diagnostic evaluations both domestically and internationally, identifying methods for increasing the cross-cultural sensitivity of this measure is essential. As insufficient norms endure as a primary criticism of ASD diagnostic instruments such as the ADOS (Matson et al. 2007), one approach may be to develop norms that can be used for different cultural contexts. The complexity of human behavior across cultures presents substantial practical barriers to developing multiple sets of norms; yet, this would represent important forward progress. The issue of non-diverse norms is a widespread issue in the ASD field and is not limited to the ADOS. Indeed, norm-based and criterion-based approaches both rely on comparisons to a majority standard.

Ethnic and racial minorities are continuously underrepresented in ASD research (Hilton et al. 2010). This sample bias is particularly concerning in light of disparities among racial/ethnic groups in the United States in terms of age and rate of ASD diagnosis and subsequent treatment quality and utilization (Magaña et al. 2013; Mandell et al. 2002). The current research adds to the body of research aiming to understand mechanisms of disparities in

ASD by highlighting a potential bias in diagnostic instruments. These findings generally speak to the need for more research assessing the reasons for identified differences in ASD symptom presentation and diagnostic rates across different racial and ethnic groups to aid in more accurate diagnosis. Forward progress in this area may help to both more clearly hone in on an etiological explanation for observed race/ethnicity-based differences in rates of ASD diagnosis and reduce disparities.

Acknowledgments Ashley J. Harrison's work on this project was supported by a NIMH T32 fellowship to (5T32MH019927_20). The authors would like to express their appreciation to the following students for their help with this project and manuscript, Margaret Naughton, Rebecca Miller, and Alejandra Perez-Ramirez and the families that participated in the Simons Simplex Collection data set.

Author Contributions AJH and RNJ conceived of the study and designed the study. AJH requested SSC data access. RNJ and DCT performed the statistical analysis and the provided the written interpretation of the data analyses. AJH and KAL helped interpret the data and aided in contextualizing the findings in the cross-cultural literature. All authors drafted and revised the manuscript.

Compliance with Ethical Standards

Conflict of interests All the authors declare that they have no conflict of interest.

Ethical Approval All procedures performed were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent Informed consent was obtained from all participants included in the study.

References

Al-Salehi, S. M., & Al-Hifthy, E. H. (2009). Autism in Saudi Arabia: Presentation, clinical correlates and comorbidity. *Transcultural Psychiatry, 46*(2), 340–347.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders (4th ed., text revision)*. Washington, DC: American Psychiatric Association.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental health disorders* (5th ed.). Washington, DC: American Psychiatric Association.

Begeer, S., El Bouk, S., Boussaid, W., Terwogt, M. M., & Koot, H. M. (2009). Underdiagnosis and referral bias of autism in ethnic minorities. *Journal of Autism and Developmental Disorders, 39*(1), 142–148. doi:10.1007/s10803-008-0611-5.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238.

Bracken, B. A., & Barona, A. (1991). State of the art procedures for translating, validating and using psychoeducational tests in cross-cultural assessment. *School Psychology International, 12*, 119–132. doi:10.1177/0143034391121010.

Browne, M. W., Cudeck, R., Bollen, K. A., & Long, J. S. (1993). Alternative ways of assessing model fit. *Sage Focus Editions, 154*, 136–136.

Caron, K. G., Schaaf, R. C., Benevides, T. W., & Gal, E. (2012). Cross-cultural comparison of sensory behaviors in children with autism. *American Journal of Occupational Therapy, 66*(5), e77–e80.

Carter, J. A., Lees, J. A., Murira, G. M., Gona, J., Neville, B. G., & Newton, C. R. (2005). Issues in the development of cross-cultural assessments of speech and language for children. *International Journal of Language & Communication Disorders, 40*(4), 385–401. doi:10.1080/13682820500057301.

Chaidez, V., Hansen, R. L., & Hertz-Picciotto, I. (2012). Autism spectrum disorders in Hispanics and non-Hispanics. *Autism: The International Journal of Research and Practice, 16*(4), 381–397. doi:10.1177/1362361311434787.

Chua, H. F., Boland, J. E., & Nisbett, R. E. (2005). Cultural variation in eye movements during scene perception. *Proceedings of the National Academy of Sciences of the United States of America, 102*(35), 12629–12633. doi:10.1073/pnas.0506162102.

Chung, K.-M., Jung, W., Yang, J.-W., Ben-Itzchak, E., Zachor, D. A., Furniss, F., … Barker, A. A. (2012). Cross cultural differences in challenging behaviors of children with autism spectrum disorders: An international examination between Israel, South Korea, the United Kingdom, and the United States of America. *Research in Autism Spectrum Disorders, 6*(2), 881–889.

Collett, P. (1971). Training Englishmen in the non-verbal behaviour of Arabs. *International Journal of Psychology, 6*(3), 209–215. doi:10.1080/00207597108246684.

Constantino, J. N., & Gruber, C. P. (2012). *Social responsiveness scale, second edition: SRS-2*. Los Angeles: Western Psychological Services.

Daley, T. C. (2004). From symptom recognition to diagnosis: children with autism in urban India. *Social Science & Medicine, 58*(7), 1323–1335.

El Bouk, S., Boussaid, W., Meerum Terwogt, M., & Koot, H. (2009). Underdiagnosis and referral bias of autism in ethnic minorities. *Journal of Autism and Developmental Disorders, 39*(1), 142–148.

Elfenbein, H. A. (2013). Nonverbal dialects and accents in facial expressions of emotion. *Emotion Review, 5*(1), 90–96. doi:10.1177/175407391245133.

Elfenbein, H. A., Beaupré, M., Lévesque, M., & Hess, U. (2007). Toward a dialect theory: Cultural differences in the expression and recognition of posed facial expressions. *Emotion, 7*(1), 131. doi:10.1037/1528-3542.7.1.131.

Ember, C. R., & Cunnar, C. M. (2015). Children's play and work: The relevance of cross-cultural ethnographic research for archaeologists. *Childhood in the Past, 8*(2), 87–103. doi:10.1179/1758571615Z.00000000031.

Farver, J. M., & Howes, C. (1993). Cultural differences in American and Mexican mother-child pretend play. *Merrill-Palmer Quarterly, 39*, 344–358.

Fischbach, G. D., & Lord, C. (2010). The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron, 68*(2), 192–195.

Freeth, M., Milne, E., Sheppard, E., & Ramachandran, R. (2014). Autism across cultures: Perspectives from non-western cultures and implications for research. In F. R. Volkmar, S. J. Rogers, R. Paul, & K. A. Pelphrey (Eds.), *Handbook of autism and pervasive developmental disorders*, (4th ed.,Vol. 2, pp. 997–1013). Hoboken, NJ: Wiley.

Freeth, M., Sheppard, E., Ramachandran, R., & Milne, E. (2013). A cross-cultural comparison of autistic traits in the UK, India and Malaysia. *Journal of Autism and Developmental Disorders, 43*(11), 2569–2583.

Fugita, S. S., Wexley, K. N., & Hillery, J. M. (1974). Black-White differences in nonverbal behavior in an interview setting. *Journal of Applied Social Psychology, 4*(4), 343–350. doi:10.1111/j.1559-1816.1974.tb02606.x.

Gotham, K., Risi, S., Dawson, G., Tager-Flusberg, H., Joseph, R., Carter, A., … Hyman, S. L. (2008). A replication of the Autism Diagnostic Observation Schedule (ADOS) revised algorithms. *Journal of the American Academy of Child & Adolescent Psychiatry, 47*(6), 642–651. doi:10.1097/CHI.0b013e31816bffb7.

Hall, W. S., & Freedle, R. O. (1975). *Culture and language: The black American experience*. Washington D.C.: Hemisphere Publishing.

Hall, W. S., Reder, S., & Cole, M. (1975). Story recall in young Black and White children: Effects of racial group membership, race of experimenter, and dialect. *Developmental Psychology, 11*(5), 628. doi:10.1037/0012-1649.11.5.628.

Hess, U., Adams, R., & Kleck, R. (2005). Who may frown and who should smile? Dominance, affiliation, and the display of happiness and anger. *Cognition & Emotion, 19*(4), 515–536. doi:10.1080/02699930441000364.

Hiller, R. M., Young, R. L., & Weber, N. (2015). Sex differences in pre-diagnosis concerns for children later diagnosed with autism spectrum disorder. *Autism: The International Journal of Research and Practice, 20*(1), 75–84. doi:10.1177/1362361314568899.

Hilton, C. L., Fitzgerald, R. T., Jackson, K. M., Maxim, R. A., Bosworth, C. C., Shattuck, P. T., … Constantino, J. N. (2010). Brief report: Under-representation of African Americans in autism genetic research: A rationale for inclusion of subjects representing diverse family structures. *Journal of Autism and Developmental Disorders, 40*(5), 633–639. doi:10.1007/s10803-009-0905-2.

Hu, L.-T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*(4), 424.

Jack, R. E., Garrod, O. G., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences, 109*(19), 7241–7244. doi:10.1097/01.mlr.0000245250.50114.0f.

Jones, R. N. (2006). Identification of measurement differences between English and Spanish language versions of the Mini-Mental State Examination: detecting differential item functioning using MIMIC modeling. *Medical Care, 44*(11), S124–S133.

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.

Knapp, M. L., Hall, J. A., & Horgan, T. G. (2013). *Nonverbal communication in human interaction*. Boston: Wadsworth Cengage Learning.

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. doi:10.1007/s10803-007-0475-0.

Kochman, T. (1981). *Black and white styles in conflict*: Chicago: University of Chicago Press.

LaFrance, M., & Hecht, M. A. (2000). Gender and smiling: A meta-analysis. In A. Fischer (Ed.), *Gender and emotion: Social psychological perspectives* (pp. 118). Cambridge, UK: Cambridge University Press.

LaFrance, M., & Mayo, C. (1976). Racial differences in gaze behavior during conversations: Two systematic observational studies. *Journal of Personality and Social Psychology, 33*(5), 547.

Lai, M.-C., Lombardo, M., Auyeung, B., Chakrabarti, B., & Baron-Cohen, S. (2015). Sex/gender differences and autism. *Journal of the American Academy of Child & Adolescent Psychiatry, 54*(1), 11–24.

Lancy, D. F. (2007). Accounting for variability in mother–child play. *American Anthropologist, 109*(2), 273–284. doi:10.1525/aa.2007.109.2.273.

Landa, R., & Garrett-Mayer, E. (2006). Development in infants with autism spectrum disorders: A prospective study. *Journal of Child Psychology & Psychiatry, 47*(6), 629–638. doi:10.1111/j.1469-7610.2006.01531.x.

Liptak, G. S., Benzoni, L. B., Mruzek, D. W., Nolan, K. W., Thingvoll, M. A., Wade, C. M., & Fryer, G. E. (2008). Disparities in diagnosis and access to health services for children with autism: Data from the National Survey of Children's Health. *Journal of Developmental & Behavioral Pediatrics, 29*(3), 152–160.

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

Lord, C., Rutter, M., DiLavore, P., & Risi, S. (2002). *Autism Diagnostic Observation Schedule: ADOS*. Los Angeles: Western Psychological Services.

Lord, C., Rutter, M., DiLavore, P., Risi, S., Gotham, K., & Bishop, S. L. (2012). *Autism Diagnsotic Observation Schedule, second edition (ADOS-2) manual (Part 1): Modules 1–4*. Torrance, CA: Western Psychological Services.

Magaña, S., Lopez, K., Aguinaga, A., & Morton, H. (2013). Access to diagnosis and treatment services among Latino children with autism spectrum disorders. *Intellectual and Developmental Disabilities, 51*(3), 141–153. doi:10.1352/1934-9556-51.3.141.

Magiati, I., Goh, D. A., Lim, S. J., Gan, D. Z. Q., Leong, J., Allison, C., … Saw, S. M. (2015). The psychometric properties of the Quantitative-Checklist for Autism in Toddlers (Q-CHAT) as a measure of autistic traits in a community sample of Singaporean infants and toddlers. *Molecular Autism, 6*(1), 40.

Mandell, D. S., Listerud, J., Levy, S. E., & Pinto-Martin, J. A. (2002). Race differences in the age at diagnosis among Medicaid-eligible children with autism. *Journal of the American Academy of Child & Adolescent Psychiatry, 41*(12), 1447–1453. doi:10.1097/01.CHI.0000024863.60748.53.

Mandell, D. S., Wiggins, L. D., Carpenter, L. A., Daniels, J., DiGuiseppi, C., Durkin, M. S., Giarelli, E., … Pinto-Martin, J. A. (2009). Racial/ethnic disparities in the identification of children with autism spectrum disorders. *American Journal of Public Health, 99*(3), 493.

Marsh, A. A., Elfenbein, H. A., & Ambady, N. (2003). Nonverbal "accents" cultural differences in facial expressions of emotion. *Psychological Science, 14*(4), 373–376.

Matson, J. L., Matheis, M., Burns, C., Esposito, G., Venuti, P., Pisula, E., … Kamio, Y. (2017). Examining cross-cultural differences in autism spectrum disorder: A multinational comparison from Greece, Italy, Japan, Poland, and the United States. *European Psychiatry, 42*, 70–76.

Matson, J. L., Nebel-Schwalm, M., & Matson, M. L. (2007). A review of methodological issues in the differential diagnosis of autism spectrum disorders in children. *Research in Autism Spectrum Disorders, 1*(1), 38–54. doi:10.1016/j.rasd.2006.07.004.

Matson, J. L., Worley, J. A., Fodstad, J. C., Chung, K.-M., Suh, D., Jhin, H. K., … Furniss, F. (2011). A multinational study examining the cross cultural differences in reported symptoms of autism spectrum disorders: Israel, South Korea, the United Kingdom, and the United States of America. *Research in Autism Spectrum Disorders, 5*(4), 1598–1604.

Mazefsky, C. A., & Oswald, D. P. (2006). The discriminative ability and diagnostic utility of the ADOS-G, ADI-R, and GARS for children in a clinical setting. *Autism: The International Journal of Research and Practice, 10*(6), 533–549.

McCarthy, A., Lee, K., Itakura, S., & Muir, D. W. (2006). Cultural display rules drive eye gaze during thinking. *Journal of Cross-Cultural Psychology, 37*(6), 717–722. doi:10.1177/0022022106292079.

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. doi:10.1037/0033-2909.126.3.424.

Mundy, P. (1995). Joint attention and social-emotional approach behavior in children with autism. *Development and Psychopathology, 7*(01), 63–82. doi:10.1017/S0954579400006349.

Muthén, L. K., & Muthén, B. O. (2005). *Mplus: Statistical analysis with latent variables: User's guide*. Los Angeles: Muthén & Muthén.

Norbury, C. F., & Sparks, A. (2013). Difference or disorder? Cultural issues in understanding neurodevelopmental disorders. *Developmental Psychology, 49*(1), 45. doi:10.1037/a0027446.

Ozonoff, S., Goodlin-Jones, B. L., & Solomon, M. (2005). Evidence-based assessment of autism spectrum disorders in children and adolescents. *Journal of Clinical Child and Adolescent Psychology, 34*(3), 523–540. doi:10.1207/s15374424jccp3403_8.

Pae, T.-I., & Park, G.-P. (2006). Examining the relationship between differential item functioning and differential test functioning. *Language Testing, 23*(4), 475–496.

Richesson, R., Anderson, M., & Smerek, M. (2014). *Race/ethnicity data standards*. Retrieved from https://www.nihcollaboratory.org/Products/RaceEthnicity_standard.pdf.

Runnels, J. (2013). Measuring differential item and test functioning across academic disciplines. *Language Testing in Asia, 3*(1), 9. doi:10.1186/2229-0443-3-9.

Satorra, A. (2000). Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In *Innovations in Multivariate Statistical Analysis*, pp. 233–247. New York: Springer.

Schofield, T. J., Parke, R. D., Kim, Y., & Coltrane, S. (2008). Bridging the acculturation gap: Parent-child relationship quality as a moderator in Mexican American families. *Developmental Psychology, 44*(4), 1190. doi:10.1037/a0012529.

Sell, N. K., Giarelli, E., Blum, N., Hanlon, A. L., & Levy, S. E. (2012). A comparison of autism spectrum disorder DSM-IV criteria and associated features among African American and white children in Philadelphia County. *Disability and Health Journal, 5*(1), 9–17. doi:10.1016/j.dhjo.2011.08.002.

Simons. (2010). *SFARI base/SSC researcher welcome packet*. Retrieved April 19, 2010, from http://simonsfoundation.s3.amazonaws.com/share/SFARI_Researcher_Welcome.pdf.

Sipes, M., Furniss, F., Matson, J. L., & Hattier, M. (2012). A multinational study examining the cross cultural differences in social skills of children with autism spectrum disorders: A comparison between the United Kingdom and the United States of America. *Journal of Developmental and Physical Disabilities, 24*(2), 145–154.

Soto, S., Linas, K., Jacobstein, D., Biel, M., Migdal, T., & Anthony, B. J. (2014). A review of cultural adaptations of screening tools for autism spectrum disorders. *Autism, 19*(6), 646–661.

United States Census Bureau. (2017). *Race*. Retrieved from https://www.census.gov/topics/population/race/about.html.

Vrana, S. R., & Rollock, D. (2002). The role of ethnicity, gender, emotional content, and contextual differences in physiological, expressive, and self-reported emotional responses to imagery. *Cognition & Emotion, 16*(1), 165–192. doi:10.1080/0269993014300018.

Western Psychological Services. (2016). *Published translations*. Retrieved from http://www.wpspublish.com/app/OtherServices/PublishedTranslations.aspx.

Yuki, M., Maddux, W. W., & Masuda, T. (2007). Are the windows to the soul the same in the East and West? Cultural differences in using the eyes and mouth as cues to recognize emotions in Japan and the United States. *Journal of Experimental Social Psychology, 43*(2), 303–311. doi:10.1016/j.jesp.2006.02.004.

Zebrowitz, L. A., Kikuchi, M., & Fellous, J.-M. (2010). Facial resemblance to emotions: Group differences, impression effects, and race stereotypes. *Journal of Personality and Social Psychology, 98*(2), 175. doi:10.1037/a0017990.