

Standardizing ADOS Scores for a Measure of Severity in Autism Spectrum Disorders

Katherine Gotham · Andrew Pickles ·
Catherine Lord

Published online: 12 December 2008
© Springer Science+Business Media, LLC 2008

Abstract The aim of this study is to standardize Autism Diagnostic Observation Schedule (ADOS) scores within a large sample to approximate an autism severity metric. Using a dataset of 1,415 individuals aged 2–16 years with autism spectrum disorders (ASD) or nonspectrum diagnoses, a subset of 1,807 assessments from 1,118 individuals with ASD were divided into narrow age and language cells. Within each cell, severity scores were based on percentiles of raw totals corresponding to each ADOS diagnostic classification. Calibrated severity scores had more uniform distributions across developmental groups and were less influenced by participant demographics than raw totals. This metric should be useful in comparing assessments across modules and time, and identifying trajectories of autism severity for clinical, genetic, and neurobiological research.

Keywords Autism spectrum disorders ·
Autism diagnostic observation schedule (ADOS) ·
Severity

Introduction

Currently, levels of impairment in children with autism spectrum disorders (ASD) are measured largely in terms of language delay, cognitive functioning, or behavioral issues

such as aggression. While these are important factors in overall adaptive functioning, they are not core features of the autism spectrum. Measuring the relative severity of autism-specific features could contribute to our ability to accurately describe ASD phenotypes across samples and across time in clinical and treatment research. An ASD severity metric could be useful for categorizing samples into more homogeneous groups in genetic and other neurobiological studies; it would also address a need to document severity as part of clinical assessment.

At this point, measures that provide autism severity ratings, such as the Childhood Autism Rating Scale (CARS; Schopler et al. 1986), the Gilliam Autism Rating Scale (GARS; Gilliam 1995), or the Autism Behavior Checklist (ABC; Krug et al. 1980), tend to yield scores that are either strongly correlated with IQ or that do not correspond to standard measures of diagnosis (Gilliam 1995; Volkmar et al. 1988; Spiker et al. 2002; South et al. 2002; Szatmari et al. 2003). The Social Responsiveness Scale (SRS; Constantino et al. 2003) provides a method for quantifying social impairment that has shown relative independence from participant characteristics such as IQ. SRS scores are based on parent or teacher report, however, and thus a complementary measure of ASD severity that offers the opportunity to take into account the observations of an experienced clinician would be desirable.

For genetic, neuroscience, and intervention research, severity of core autism features often has been estimated using primary phenotyping measures, the Autism Diagnostic Observation Schedule (ADOS; Lord et al. 2000) and the Autism Diagnostic Interview-Revised (ADI-R; Rutter et al. 2003). While it is true that higher ADI-R and ADOS scores indicate that an individual has a greater number of items representing core deficits and/or greater severity of impairment, scores were not normalized for this purpose

K. Gotham (✉) · C. Lord
University of Michigan Autism and Communication Disorders
Center, 1111 East Catherine Street, Ann Arbor, MI 48109-2054,
USA
e-mail: kog@umich.edu

A. Pickles
University of Manchester, Manchester, UK

and vary in the degree to which they are correlated with both IQ and chronological age. Attempts to indicate severity using ADI-R item scores selected to operationalize ICD-10 criteria for the disorder proved successful in predicting the number of affected relatives of verbal probands, but not for nonverbal probands (Pickles et al. 2000). One limitation of ADI-R scores as a severity metric is that nonverbal children are not scored on roughly 25% of the total ADI-R items, and so communication domain summary scores are restricted by non-random missing data.

The ADOS, a semi-structured autism diagnostic observation, has shown strong predictive validity against best estimate diagnoses (Gotham et al. 2007), making it a common choice among phenotyping measures. In each of four developmental- and language-level dependent modules, a protocol of social presses is administered by a trained examiner, and then behavioral items relevant to ASD are scored on a 4-point scale, with 0 indicating 'no abnormality of type specified' and 3 indicating 'moderate to severe abnormality.' Specific items comprise an algorithm for each module; these items are summed and compared to thresholds, which results in a classification of "autism," "autism spectrum disorder," or "nonspectrum."

Because the ADOS has been used to catalogue ASD features in large samples, ADOS raw totals are a common stand-in for a measure of autism severity. This instrument was created for diagnostic purposes, and thus was not specifically designed to facilitate longitudinal and cross-sectional comparison of data. As an individual gains language skills, he or she potentially moves through ADOS modules, making raw scores not directly comparable across time. Additionally, effects of age and language level on domain total and algorithm scores have been observed (Joseph et al. 2002; de Bildt et al. 2004; Gotham et al. 2007).

In 2007, the original ADOS algorithms were revised in part for the purpose of increasing the comparability across modules 1–3. Algorithms with the same number of items and of similar content across modules were created (Gotham et al. 2007). These revisions resulted in improved specificity of the measure among more impaired populations, while generally maintaining or improving predictive validity among individuals of other developmental levels (e.g., fluent speakers). The algorithm domain structure now includes a Social Affect (SA) and a Restricted, Repetitive Behavior (RRB) domain for each of the five developmentally-based algorithms corresponding to modules 1–3. Comparability of item content and total item number across these algorithms was intended to improve the interpretability of longitudinal comparisons using the measure. Still, items are necessarily developmentally graded across modules, making calibration necessary to compare algorithm totals.

Some effects of participant characteristics still exist within and across ADOS modules as well. Revised

algorithm totals met the goal of independence from chronological age and decreased association with verbal IQ, with the exception of Module 1 scores (Gotham et al. 2007). A replication of the algorithm revisions in an independent dataset again found low correlations between raw scores and age, verbal IQ, and nonverbal IQ, though significant associations remained between verbal IQ and SA domain total scores for Module 1 recipients with few or no single words and Module 2 recipients aged 5 or older (Gotham et al. 2008).

True normalization of severity of autism would require a representative population, but to date, population studies have been too small, e.g., Brick Township (Bertrand et al. 2001), have not used the ADOS (Chakrabarti and Fombonne 2005; CDC 2007), or have collected samples older than most clinically assessed children (Baird et al. 2006). Acknowledging these limitations, in the present study we elected to standardize ADOS scores using a large "convenience" sample of individuals with ASD. Our goals were to reduce remaining participant demographic effects to the greatest possible degree, and generate standard scores that would approximate a severity metric for the construct of 'autism spectrum' as it is measured on the ADOS. This metric ideally will be useful in (1) allowing comparison of assessments across modules and time; (2) providing a means of assessing the relationship between severity in ASD and verbal and nonverbal IQ; and (3) identifying different trajectories of autism severity independent of verbal IQ both for clinical purposes and for phenotypic subgrouping in genetic and neurobiological research. We hope that calibrated severity scores can then be replicated in smaller population-based studies and tested for validity in predicting treatment responsiveness and other clinical outcomes in children with ASD.

Our first approach to developing a severity metric was to calibrate ADOS algorithm totals using eight age/language cells chosen on the basis of theoretically-driven expectations for specific age ranges with similar developmental impairments. This would have allowed a 'prefix' on the severity score that indicated age and language level out of the eight possible groups (ranging from young Module 1's with no words to fluent speakers, aged 5–10). Within each cell, raw totals were converted to Z-scores, which were then converted to a 100-point scale. This method yielded calibrated scores that fanned out, with increasing variability of individuals' ADOS totals over time and age. Thus, an alternative approach was chosen in which a greater number of age/language cells were used, and severity scores within each cell were based on the raw total percentiles that corresponded to each of three possible ADOS diagnostic classifications. This method is described in more detail below.

Methods

Participants

Analyses were conducted on data from 1,415 individuals, of which 355 individuals with ASD diagnoses had repeated measure data. The final dataset included 2,195 assessments, where ‘assessment’ is defined as contemporaneous ADOS data and a best estimate clinical diagnosis. Autism diagnoses were assigned to 1,187 assessments (54% of entire sample); 599 assessments were given diagnoses of non-autism ASD (27% of the sample, including $n = 12$ with Asperger Disorder, $n = 3$ with Childhood Disintegrative Disorder, and $n = 584$ with Pervasive Developmental Disorder, Not Otherwise Specified, or PDD-NOS), and 409 had non-ASD developmental delays (19%). Contemporaneous verbal IQ data was available for 2,007 assessments (91.4% of the entire sample) and nonverbal IQ data for 1,989 assessments (91.0%). Please refer to Table 1 for a detailed description of the dataset by revised algorithm group.

Chronological ages in the sample ranged from 2 to 16 years (see Table 1 for age range by algorithm group). Recipients of ADOS Module 4 (older adolescents and adults with fluent speech) were not included in these analyses because of smaller sample size and the different relevance of age equivalents in adults. Females comprised 22% of the dataset ($N = 478$ assessments). Ethnicities represented by these data include 14% African American ($N = 306$ assessments); 3% Asian American ($N = 58$); 77% Caucasian ($N = 1699$); 0.5% Native American ($N = 10$); 2% biracial ($N = 40$); and other ($N = 20$) or race not specified ($N = 62$) totaling 4% of assessments. Twenty-three percent of the sample reported maternal education at the graduate or professional level; 56% of mothers had a bachelor’s degree or some college education, and 21% of mothers had a high school degree or less.

Within the nonspectrum sample of 409 assessments, 111 had a primary diagnosis of a language disorder (27% of nonspectrum total), 80 were assessments with nonspecific intellectual disability (20%), 56 with Down syndrome (14%), 55 with oppositional defiant disorder, ADD and/or ADHD (13%), 31 with mood and/or anxiety disorders (8%), 29 with Fetal Alcohol Spectrum Disorders (7%), 24 with non-ASD genetic and/or physical disabilities such as Fragile X, Williams syndrome, or mild cerebral palsy (6%), and 23 had an early delay that clinicians were not comfortable categorizing (5%).

The majority of participants were self-, school-, or physician-referred clinic patients at the University of Michigan Autism and Communication Disorders Center (UMACC) or the University of Chicago Developmental Disorders Clinic. The rest participated in a longitudinal

study conducted through the Treatment and Education of Autistic and Communication Handicapped Children (TEACCH) Centers at the University of North Carolina, Chapel Hill, and the University of Chicago clinic, or received diagnostic evaluations through recent, ongoing studies at UMACC, including those focused on participants with non-ASD developmental delays, ASD-affected sibling pairs, or children between 12 and 36 months of age who failed a social-communication screener. Out of 399 participants with repeated assessments through clinic reevaluations or longitudinal research, 301 individuals had 2 or 3 ADOS assessments (57% with autism, 31% with PDD-NOS, and 12% NS), and 98 individuals had between 4 and 8 assessments (58% with autism, 33% with PDD-NOS, and 9% NS). Individuals with longitudinal data did not differ significantly in gender, race, or maternal education from those with only one assessment point, however they had significantly lower mean verbal IQs ($M = 49.6$, $SD = 27.8$) and nonverbal IQs ($M = 73.0$, $SD = 23.8$) at first assessment than did single assessments (verbal IQ $M = 68.2$, $SD = 32.8$; nonverbal IQ $M = 77.9$, $SD = 27.5$); verbal IQ $t(1,351) = 9.7$, $p < 0.001$ and nonverbal IQ $t(1,334) = 3.0$, $p < .01$.

Measures and Procedure

The most typical research protocol across sites and projects was the initial administration of the ADI-R and the Vineland Adaptive Behavior Scales, 1st (VABS; Sparrow et al. 1984) or 2nd edition (Vineland II; Sparrow et al. 2005), to a parent or caregiver, followed by a child evaluation in which psychometric testing preceded the ADOS. The second most common protocol was a re-evaluation consisting of psychometric testing and an ADOS. In both cases, a clinical diagnosis was made by a psychologist and/or psychiatrist after review of all data. The ADI-R was available for 1,700 assessments (77% of sample) and the Vineland for 1,710 assessments (78%). The ADOS was administered and scored by a clinical psychologist or trainee who met standard requirements for research reliability. The Pre-Linguistic Autism Diagnostic Observation Schedule (PL-ADOS; DiLavore et al. 1995) was given in 418 assessments (19%) and the piloted ADOS-T (Luyster et al. 2009), a toddler version of the ADOS, was given in 82 assessments (4%); for both measures, identical items were recorded to Module 1 algorithm scores. A developmental hierarchy of cognitive measures, most frequently the Mullen Scales of Early Learning (MSEL; Mullen 1995) and the Differential Ability Scales (DAS; Elliot 1990), determined IQ scores.

Clinic-referred participants received oral feedback and a written report without financial compensation. Participants

Table 1 Sample description

DX	Module 1, No Words			Module 1, Some Words			Module 2, Younger than 5			Module 2, 5 or Older			Module 3							
	N	Mean	SD	Range	N	Mean	SD	Range	N	Mean	SD	Range	N	Mean	SD	Range				
Autism																				
Age	475	52.59	27.49	24–174	281	56.63	24.99	25–162	89	48.55	7.34	27–59	164	98.46	32.03	60–196	178	103.23	29.35	48–185
viq	448	25.38	14.03	2–92	254	46.66	19.69	10–137	81	77.00	20.73	28–132	155	52.64	19.33	20–107	156	87.56	23.22	31–155
nviq	442	52.34	21.20	2–144	253	65.82	20.54	14–122	77	90.44	22.87	52–170	155	77.25	23.98	24–131	158	92.98	23.57	34–155
vma	453	11.39	19.05	1–291	251	25.29	30.59	0–353	82	42.94	50.09	21–357	157	54.91	53.34	13–354	151	87.50	44.66	3–369
nvma	443	24.33	10.31	5–96	254	35.10	15.04	13–110	78	46.94	18.45	17–152	146	71.84	29.42	28–194	149	93.70	31.55	25–167
ADI social	419	21.28	530	2–30	212	19.21	6.43	0–30	68	16.12	4.83	3–27	107	22.32	5.90	0–30	116	19.55	5.21	6–30
ADI comm-V	5	17.40	4.09	12–22	74	15.05	4.50	0–24	64	15.41	3.99	3–22	101	18.18	4.11	0–26	116	16.99	4.30	7–25
ADI comm-NV	419	11.74	2.34	0–14	212	9.89	3.35	0–14	68	8.57	2.93	0–14	107	10.36	3.25	0–14	116	9.27	3.27	2–14
ADI-RR	419	4.79	1.72	0–10	212	5.41	2.20	0–10	68	6.34	2.59	1–12	107	7.04	2.67	0–12	116	7.50	2.72	2–12
ADOS SA	475	17.29	2.34	5–20	281	14.87	3.38	1–20	89	12.93	3.45	2–20	164	14.47	3.39	6–20	178	11.39	3.87	0–20
ADOS RR	475	4.92	2.02	0–8	281	4.43	2.00	0–8	89	4.51	1.98	0–8	164	5.12	1.95	0–8	178	3.44	1.94	0–8
PDD-NOS																				
Age	76	38.03	13.48	24–107	114	43.84	17.78	24–175	108	42.77	9.98	24–59	51	78.96	18.33	60–129	250	101.98	31.26	42–195
viq	74	35.68	15.67	7–75	107	66.78	20.02	11–129	83	84.50	20.57	2–149	44	63.93	17.56	28–102	230	100.93	21.19	49–164
nviq	73	58.09	22.04	15–99	105	79.19	21.49	24–133	84	94.44	22.72	2–159	46	74.39	21.55	31–118	225	98.48	21.16	14–153
vma	75	12.96	6.69	2–47	104	33.80	46.33	11–357	81	64.04	88.86	18–354	45	57.93	61.65	26–371	226	110.88	73.26	39–498
nvma	73	20.34	6.72	7–34	105	40.10	47.87	16–507	80	41.66	13.30	23–91	44	59.07	13.05	34–90	208	102.30	36.22	37–190
ADI social	72	15.15	5.67	3–25	94	11.57	6.49	1–28	65	10.72	5.11	2–25	33	14.82	7.23	4–29	174	14.93	7.43	0–29
ADI comm-V	0	–	–	–	58	10.41	4.97	2–25	53	11.19	4.64	0–21	32	12.44	4.04	3–19	174	12.75	5.40	0–24
ADI comm-NV	72	9.94	2.95	2–14	94	6.26	3.97	0–14	64	5.88	3.58	0–14	33	6.03	3.44	0–12	174	6.59	3.80	0–14
ADI-RR	72	3.56	2.23	0–9	94	3.94	2.38	0–10	65	4.45	2.62	0–10	33	4.73	3.01	1–12	174	5.39	3.06	0–12
ADOS SA	76	13.69	4.24	1–20	114	9.17	4.11	0–19	108	8.37	3.81	0–18	51	9.11	4.36	0–19	250	7.73	4.06	0–19
ADOS RR	76	3.10	1.94	0–8	114	3.17	1.96	0–8	108	3.39	2.07	0–8	51	3.27	1.90	0–8	250	2.19	1.62	0–7
Non-spectrum																				
age	60	39.55	19.33	24–129	107	42.10	19.23	24–129	57	43.98	7.38	27–59	44	95.95	30.64	61–184	141	107.35	29.59	50–192
viq	57	40.96	18.72	14–83	90	68.07	23.74	11–117	51	85.33	21.83	2–140	44	58.09	19.06	17–103	135	91.69	22.28	26–163
nviq	55	58.80	28.72	13–132	89	70.52	23.74	15–116	49	92.04	20.46	2–133	44	61.93	24.13	24–118	136	89.85	22.23	35–151
vma	57	13.77	5.63	1–26	87	27.62	8.34	13–52	50	56.62	75.44	17–356	43	50.14	13.94	20–77	134	103.18	61.51	32–492
nvma	55	20.62	8.67	4–47	86	29.29	8.55	15–52	46	41.67	9.90	19–73	44	56.59	16.65	26–93	132	95.02	32.57	34–189
ADI social	51	11.33	7.09	0–26	78	7.63	5.60	0–24	45	9.67	5.12	1–21	36	13.19	6.58	4–28	130	9.24	6.81	0–24
ADI comm-V	0	–	–	–	38	5.53	3.32	0–13	43	9.70	5.04	1–21	34	11.00	5.09	2–23	130	8.27	5.46	0–24
ADI comm-NV	51	7.73	4.22	0–14	78	4.29	3.26	0–12	45	5.22	3.68	0–14	36	6.17	4.08	0–14	130	4.32	3.72	0–14
ADI-RR	51	2.59	1.94	0–8	78	2.32	1.73	0–8	45	4.04	3.31	0–11	36	4.25	2.54	0–10	130	3.65	2.69	0–11

Table 1 continued

DX	Module 1, No Words			Module 1, Some Words			Module 2, Younger than 5			Module 2, 5 or Older			Module 3							
	N	Mean	SD	Range	N	Mean	SD	Range	N	Mean	SD	Range	N	Mean	SD	Range				
ADOS SA	60	8.36	5.82	0–20	107	4.71	3.91	0–17	57	3.56	2.77	0–15	44	4.16	3.14	0–10	141	3.90	2.94	0–14
ADOS RR	60	1.88	1.87	0–7	107	1.40	1.49	0–7	57	1.49	1.42	0–5	44	1.63	1.64	0–5	141	0.98	1.15	0–5

Note: All ages in months *vig* verbal IQ, *nvig* nonverbal IQ, *vma* verbal mental age, *nvma* nonverbal mental age, *ADI social* ADI-R social total, *ADI-R comm-V* ADI-R communication total for verbal subjects, *ADI-R comm-NV* ADI-R communication total for nonverbal subjects, *ADI-RR* ADI-R restricted, repetitive behaviors total, *ADOS SA* revised algorithm social affect domain, *ADOS RR* revised algorithm restricted, repetitive behavior domain

recruited only for the purpose of research received financial compensation and a written summary of evaluation results. Institutional Review Boards at the University of Chicago or the University of Michigan approved all procedures related to this project.

Mapping a Standardized Severity Metric onto Raw ADOS Scores

Severity scores were created by dividing the pool of assessments from individuals with ASD into narrowly defined age and language cells, and standardizing raw total scores from the revised algorithms (Gotham et al. 2007) within these cells. In order to maximize the number of cases available for standardization, assessments missing data from any one item from either the SA or RRB domains of the revised ADOS algorithms were retained by adding to the domain total an average item score from that participant’s existing domain data. The ASD sample alone was used for raw total standardization: this included all assessments corresponding to a best estimate diagnosis of autism or ASD, as well as data from 13 individuals who had ADOS data with a contemporaneous nonspectrum diagnosis but who were later diagnosed with ASD. This subsample ($N = 1,807$ assessments from 1,118 individuals) was separated into groups based on the five revised algorithms used with children: Module 1 No Words, Module 1 Some Words, Module 2 Younger than 5; Module 2 Age 5 and Older; and Module 3. Within each of these five developmental cells, distributions of summed SA and RRB totals were generated separately for every 1-year age group between 2 and 16 years; these age cells were collapsed when possible in order to create the fewest number of age- and language-level-determined ‘calibration cells’ with similar raw total score distributions. Younger age cells were purposely kept distinct to anticipate developmental changes and more frequent assessments in young children as they transition from toddlerhood to preschool to school programs. Age cells with similar distributions were collapsed only within the same algorithm. Eighteen calibration cells resulted (see Fig. 1).

Within each of these 18 cells, raw ADOS totals were mapped onto a 10-point severity metric. After considering a variety of approaches, severity scores 1–3 were set so as to represent the distribution of raw scores receiving a nonspectrum ADOS classification within that calibration cell, severity scores 4–5 represented ASD-classification ADOS totals, and 6–10 represented raw totals receiving an autism classification within that cell. ADOS classification thresholds were determined by the revised algorithm relevant to each calibration cell. The range of raw totals corresponding to each point on the severity metric was determined by the percentiles of available data associated

Fig. 1 Age by language level calibration cells. *Note.* N's denote the number of ASD assessments within each cell

AGE (in years)		2	3	4	5	6	7	8	9	10+
Module 1	No Words	n=203	n=141	n=130		n=86				
	Single Words	n=96	n=118	n=82	n=68		n=40			
Module 2	Phrases	n=43	n=63	n=94	n=103		n=53		n=59	
Module 3	Fluent	n=71				n=236				n=121

with each severity point within a classification range. Lower severity scores are associated with less autism impairment. Table 2 shows the raw score range corresponding to each severity point within each calibration cell.

Design and Analysis

Distributions of raw totals and severity scores were compared to assess whether severity score distributions across age/language cells were more uniform than raw score distributions. Linear regression models were analyzed to compare the relative independence of severity scores and raw totals from participant characteristics, such as chronological age, verbal and nonverbal IQ, and verbal and nonverbal “current” mental ages. Several assessments with longitudinal data were then chosen to exemplify various patterns of severity change over time across diagnostic groups.

Results

Comparing Distributions of Severity Scores and Raw Ados Totals by Calibration Cell

In line with the goal of increasing comparability across modules and developmental levels, severity scores for ASD participants were expected to have a more uniform distribution across age- and language-level calibration cells than would raw totals. Distributions of raw ADOS totals were generated for each of the 18 calibration cells (Fig. 2) and compared to the distribution of severity scores for each cell (Fig. 3).

Distributions of severity scores showed increased comparability across the age/language cells, though they were not uniform. The means and standard deviations of both severity scores and raw totals are listed by age/language cell in Table 3.

Severity score distributions exhibited a ceiling effect that is inherent to the metric. By ensuring that scores 6–10 correspond to approximate fifths of the ASD participants who received scores in the autism classification range, roughly 20% of participants received the maximum score

of ‘10’ (in this dataset, 19.3% of participants with an autism classification on the ADOS have a severity score of ‘10,’ which is 16.5% of all participants). Though some overlap exists, severity scores showed expected heterogeneity of distribution across the three diagnostic groups: autism, PDD-NOS, and nonspectrum (see Fig. 4).

Relative Independence of Severity Score from Participant Characteristics

Multiple linear regression analyses were performed separately for the dependent variables *severity score* and *raw total* to examine whether participant characteristics such as age and IQ would be less associated with severity scores than they were with raw scores. For *ASD assessments* with complete contemporaneous demographic data ($N = 1,369$), potential predictors were entered into a structured hierarchical model, in which Block 1 included verbal and nonverbal IQ and mental age variables (which are known to affect the expression of ASD symptoms; Lord and Spence 2006), and Block 2 included age, gender, maternal education, and race (variables that could affect ASD symptoms but that often have had non-significant effects when Block 1 variables are controlled). Whereas 44% of the variance in raw totals was explained by this model, only 12% of variance was explained for severity scores using these covariates. Verbal IQ and one maternal education variable (mothers with graduate/professional degrees vs. all others) emerged as significant predictors for both severity score and raw score. Nonverbal IQ, verbal mental age, nonverbal mental age, chronological age, and gender were not significant predictors of either severity scores or raw totals for ASD participants. When covarying for these variables, as well as verbal IQ and maternal education, there was a trend for African American participants to have lower severity scores than other racial groups combined ($B = -0.35$; $\beta = -0.06$, $p = 0.04$), but this is not easily interpreted due to the confounding effects of possible referral bias. For all ASD assessments with racial affiliation data ($N = 1,749$), mean severity score for African-American participants was 7.4 ($SD = 1.8$) compared to 7.3 ($SD = 2.2$) for the combined other participant groups, $t(1,747) = -0.71$; $p = 0.48$.

Table 2 Mapping of ADOS raw totals onto calibrated severity scores

ADOS class-ification	Calibrated severity score	Raw ADOS totals																			
		Module 1, No Words					Module 1, Single Words					Module 2, phrases					Module 3, fluent				
		2 years	3 years	4-5 years	6-14 years	7-14 years	2 years	3 years	4 years	5-6 years	7-14 years	2 years	3 years	4 years	5-6 years	7-8 years	9-16 years	2-5 years	6-9 years	10-16 years	
NS	1	0-6	0-6	0-3	0-3	0-2	0-2	0-2	0-2	0-2	0-2	0-2	0-3	0-3	0-2	0-2	0-3	0-3	0-2	0-3	
	2	7-8	7-8	4-6	4-6	3-4	3-4	3-4	3-4	3-5	3-5	4-5	4-5	4-5	3-5	3-5	4	3-4	4	4	
	3	9-10	9-10	7-10	7-10	5-7	5-7	5-7	5-7	6-7	6	6	6	6-7	6-7	6-7	5-6	5-6	5-6	5-6	
ASD	4	11-13	11-14	11-12	11-13	8-10	8-9	8-9	8-10	8-9	7-8	7-8	7	8	8	8	7	7	7	7	
	5	14-15	15	13-15	14-15	11	10-11	10-11	11	10-11	9	9	8-9	-	-	-	8	8	8	8	
AUT	6	16-19	16-20	16-19	16-19	12-13	12-14	12-15	12-16	12-18	10-11	10-12	10-13	9-14	9-14	9-14	9-11	9-10	9-10	9-10	
	7	20-21	21-22	20-21	20-22	14-16	15-17	16-18	17-19	19-20	12	13-14	14-16	15-16	15-17	15-17	12	11-12	11-12	11-12	
	8	22	23	22-23	23-24	17-19	18-19	19-20	20-21	21	13-14	15-16	17-18	17-20	18-21	18-20	13-15	13-14	13-14	13-14	
	9	23-24	24	24-25	25	20-21	20-21	21-22	22-23	22-23	15-17	17-18	19-20	21-22	22-23	21-23	16-17	15-17	15-17	15-17	
	10	25-28	25-28	26-28	26-28	22-28	22-28	23-28	24-28	24-28	18-28	19-28	21-28	23-28	24-28	24-28	18-28	18-28	18-28	18-28	

Note: To derive an ADOS calibrated severity score from a raw total, clinicians should first identify the relevant column from Table 2 based on the examinee's ADOS module/revised algorithm and chronological age within that module/algorithm group. The examinee's raw ADOS total is then located within the relevant column. The corresponding Calibrated Severity Score is the number in the second column from the left that falls within the same row as the examinee's raw total. It is worth noting that Calibrated Severity Scores are assigned even to those raw totals that do not meet classification thresholds of ASD or Autism on the ADOS, since clinical judgment can overrule the measure classification and result in a spectrum diagnosis

Fig. 2 Distributions of ADOS raw total scores by age/language cells (ASD assessments only). *Note.* See Table 3 for a key of age/language cells 1–18

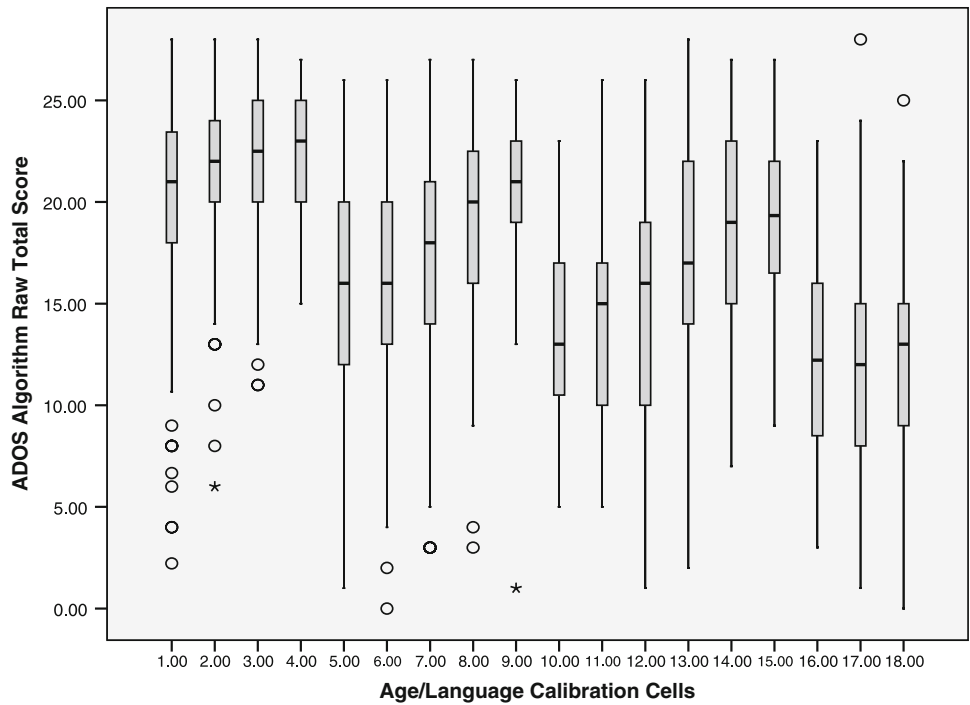
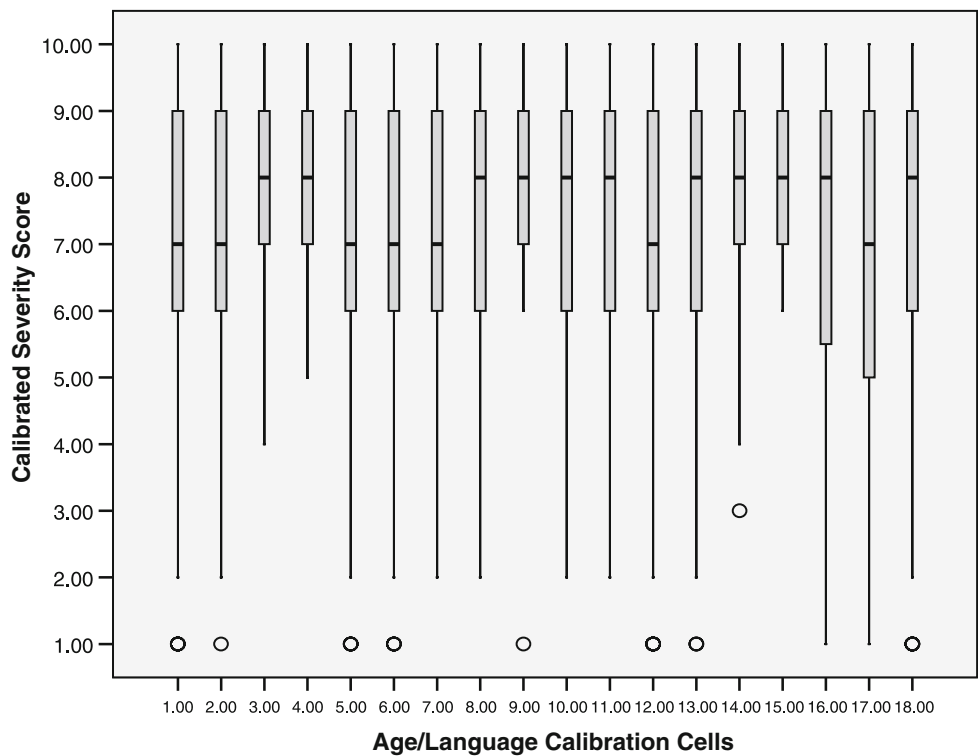


Fig. 3 Distributions of calibrated severity scores by age/language cells (ASD assessments only). *Note.* See Table 3 for a key of age/language cells 1–18



Verbal IQ and the graduate/professional maternal education variable were then entered into Forward Stepwise models (see Table 4), at which point maternal education was excluded from the model as a predictor of severity score, though retained as a predictor of raw score. Standardization reduced the effect of verbal IQ, the most

influential participant characteristic on ADOS scores. Verbal IQ explained 43% of the variance in raw totals in the model, but accounted for only 10% of the variance in severity scores in this model. This represents a change from a large effect size ($R = 0.67$) for verbal IQ on ADOS scores to an effect size just outside the accepted range for

Table 3 Raw score and calibrated severity score means and standard deviations by age/language cell (ASD assessments only)

Group	Age/Language Cell	Algorithm raw total score			Calibrated severity scores		
		N	M	SD	N	M	SD
1	Mod 1, NW, age 2	203	20.13	4.83	203	7.29	2.11
2	Mod 1, NW, age 3	141	21.63	3.85	141	7.56	1.85
3	Mod 1, NW, ages 4–5	130	21.96	3.63	130	7.87	1.48
4	Mod 1, NW, ages 6–14	86	22.35	3.34	86	7.88	1.45
5	Mod 1, SW, age 2	96	15.64	5.77	96	7.02	2.45
6	Mod 1, SW, age 3	118	15.85	5.37	118	6.99	2.26
7	Mod 1, SW, age 4	82	17.13	5.95	82	7.21	2.16
8	Mod 1, SW, ages 5–6	68	18.84	4.71	68	7.48	1.72
9	Mod 1, SW, ages 7–14	40	20.68	4.24	40	7.97	1.77
10	Mod 2, phrases, age 2	43	13.27	4.14	43	7.37	2.08
11	Mod 2, phrases, age 3	63	14.57	5.01	63	7.38	2.04
12	Mod 2, phrases, age 4	94	14.43	5.93	94	6.73	2.44
13	Mod 2, phrases, ages 5–6	103	16.84	5.78	103	7.45	1.99
14	Mod 2, phrases, ages 7–8	53	18.49	5.22	53	7.79	1.71
15	Mod 2, phrases, ages 9–16	59	19.16	4.48	59	8.10	1.37
16	Mod 3, fluent, ages 2–5	71	12.16	4.87	71	6.80	2.59
17	Mod 3, fluent, ages 6–9	236	11.66	5.19	236	6.64	2.55
18	Mod 3, fluent, ages 10–16	121	12.48	4.94	121	7.09	2.45

Note: Mod 1, NW ADOS Module 1, No Words algorithm; Mod 1, SW ADOS Module 1, Some Words Algorithm

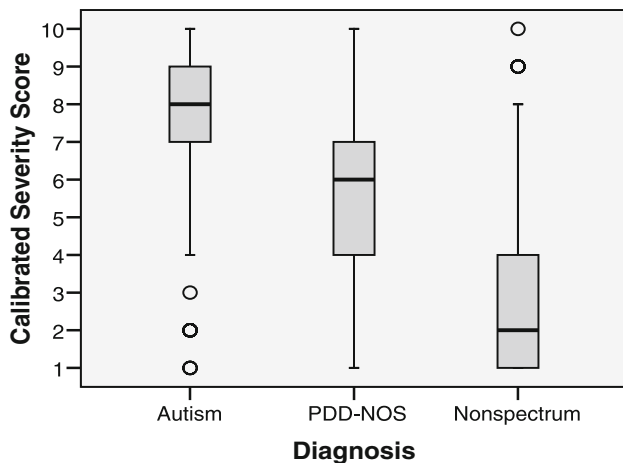


Fig. 4 Distributions of calibrated severity scores by diagnostic group

‘small’ ($R = 0.32$; see McCarthy et al. 1991; Cohen 1988). The effect of maternal education on raw total scores was likely an artifact of recruitment biases (Graduate/Professional raw total $M = 14.9$, $SD = 7.2$; other maternal education levels raw total $M = 15.4$, $SD = 7.2$; $t(1,887) = 1.13$, $p = 0.26$).

When the initial hierarchical block models were applied to the full sample (ASD and nonspectrum assessments combined), significant predictors of severity scores included verbal IQ, gender (with males the more severe group),

Table 4 Multiple linear regression models for calibrated severity scores and ADOS raw totals in ASD assessments

	R^2	F change	df	B	SE B	β
DV = severity score (ASD only, $N = 1,465$)						
Step 1 ^a	0.10	164.78	1,1463			
Constant*				8.5	.11	
Verbal IQ*				-0.02	.001	-0.32
DV = raw total (ASD only, $N = 1,465$)						
Step 1	0.43	1101.66	1,1463			
Constant*				24.14	0.24	
Verbal IQ*				-0.12	0.004	-0.66
Step 2 ^b	0.44	10.42	1,1462			
Constant*				24.05	0.24	
Verbal IQ*				-0.12	0.004	-0.67
Mat Ed*				0.94	0.29	0.07

Note: DV Dependent variable, Mat Ed Dummy coded variable separating mothers with graduate or professional education to those of all other educational levels

^a All other variables excluded from the stepwise forward model

^b Change in $R^2 = .004$ for Step 2 ($p < 0.001$)

* $p < 0.001$

and maternal education; significant predictors of raw totals included verbal IQ, nonverbal mental age, gender, chronological age, and maternal education (these statistics are available from the authors). This again indicates that, when

severity scores are applied to a clinical referral population, they are less influenced by participant characteristics than are raw ADOS totals.

Case Summaries

Four children with ASD diagnoses and longitudinal data were chosen to exemplify patterns in severity score change over time. Their scores by chronological age are plotted in Fig. 5, with ADOS module and raw total score displayed for each time point.

Case 1. “Adam,” a Caucasian male, was seen at 45 months of age as part of a clinical research project. He received a diagnosis of autism at that time. He was evaluated with ADOS Module 2 until age 13, when he received Module 3. His mental ages were 34 months nonverbal and 21 months verbal at first assessment, and 165 months nonverbal and 111 months verbal at final assessment at age 13 (NVIQ: 71 at first, 107 at last; VIQ: 44 first, 80 last). Despite his increase in IQ, Adam showed a persistently severe trajectory, with scores varying between 8 and 10 over seven assessments.

Case 2. “Bianca,” a Caucasian female, was first seen at age 48 months as a clinical referral, at which point she received a diagnosis of autism. She was evaluated with ADOS Module 2 until age 5, when she received Module 3. Her mental ages were 46 months nonverbal and 56 months

verbal at first assessment, and 107 months nonverbal and 120 months verbal at her 8.5-year-old assessment (NVIQ: 80 at first, 107 last; VIQ: 108 first, 126 last). Bianca showed decreasing autism severity over time, with scores dropping from 9 to 4 across six assessments.

Case 3. “Cara,” an African American female, was first seen as part of a research project at age 3. She received a diagnosis of autism. She was evaluated consistently using ADOS Module 1. Her mental ages were 16 months nonverbal and 8 months verbal at first assessment, and 51 months nonverbal and 11 months verbal at her last assessment at age 10 (NVIQ: 47 at first, 40 last; VIQ: 23 first, 20 last). Despite the stability of her IQ scores over time, Cara showed worsening autism severity, with scores increasing from 5 to 10 over four assessments.

Case 4. “Daniel,” a Caucasian male, was first seen at 34 months of age as a clinical referral and was given a nonspectrum diagnosis; at 46 months of age he received a PDD-NOS diagnosis which then remained stable over time. He was evaluated with ADOS Module 1 in his assessments through age 5; at age 10 he received Module 3. His mental ages were 38 months nonverbal and 36 months verbal at first assessment, and 162 months nonverbal and 142 months verbal at final assessment at age 10 (NVIQ: 112 at first, 129 at last; VIQ: 105 first, 113 last). Daniel showed consistently mild severity scores varying between 1 and 3 over four assessments.

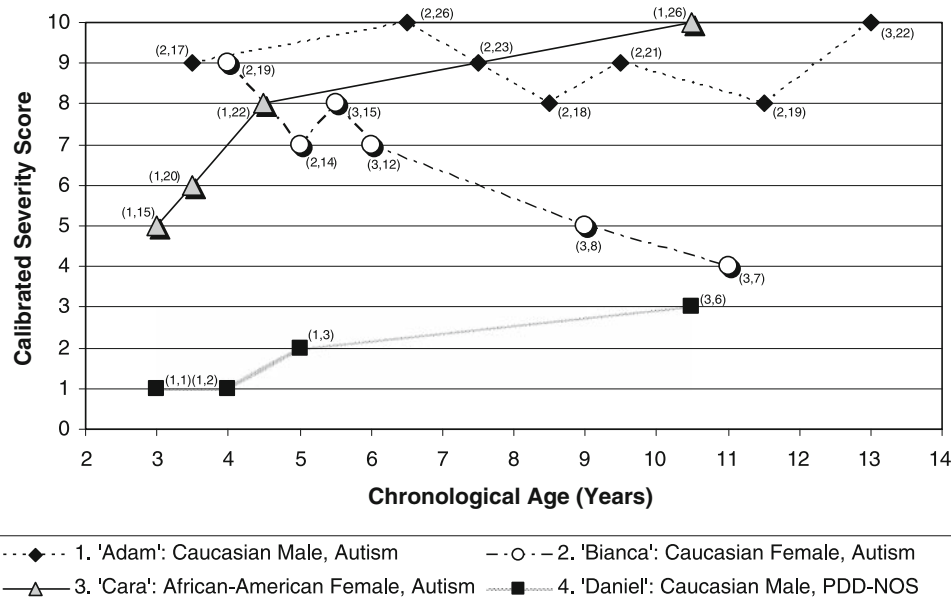


Fig. 5 Case summaries of longitudinal severity scores. *Note.* Parentheses by individual data points indicate (Module, Raw Score) for each assessment. The calibrated severity metric allows change across time and module to be evaluated in a standardized fashion in children of varying age and verbal ability. Adam and Daniel follow relatively consistent trajectories despite module changes, while a marked change in severity is apparent in Cara’s scores despite seemingly

small increases in raw total within the same module. Bianca’s decreasing raw totals alone indicate a drop in ASD severity, but the clinical import of this is obscured by her module change and increasing chronological age. Severity scores are not necessarily more stable than raw totals, but were created to allow the change or consistency in these cases to be interpreted more readily than perceived patterns in raw total scores

Discussion

The calibrated severity metric based on ADOS raw totals offers a method of quantifying ASD severity with relative independence from individual characteristics such as age and verbal IQ. It should have utility in various genetic, neurobiological, and clinical research endeavors, including treatment trials, that otherwise would use unstandardized ADOS raw totals. Calibrated scores have more uniform distributions across age- and language-groups compared to raw totals, making it possible to compare children's scores longitudinally across distinct algorithms. In part because of the modular system of the ADOS, chronological age, nonverbal IQ, and verbal and nonverbal mental age did not predict either raw totals or severity scores in this sample. The severity metric builds on this modular system to reduce the influence of participants' verbal IQ, which accounted for 10% of the variance in severity scores versus 43% of the variance in raw totals, a reduction from a large to medium effect size. The remaining influence of verbal IQ on the severity metric can be seen in the drift of mean scores toward greater severity in older age groups with lower language levels (Modules 1 and 2). This apparent age effect seems likely to be explained by lower verbal IQ in the older children without fluent speech. Though this effect has not been eliminated entirely, the calibrated metric is better able to measure autism severity beyond verbal impairment than are raw ADOS totals.

Calibrating scores within narrowly-defined age/language cells achieved the reduction in verbal IQ effects within the new metric and corrected for artificial variability in individuals' scores across time. Unfortunately, a greater number of calibration cells precludes a user-friendly age/language 'prefix' to the severity score, as mentioned in the "Introduction". The method described here necessarily defines autism severity in relation to individuals of similar age and language ability. When using these scores clinically and for research, one must keep in mind the age/language level of the child/sample, as there clearly will be developmental and adaptive functioning differences among children with the same severity score on this 10-point scale. This is true of all standardized scores. Calibrated severity scores do not measure functional impairment, but are intended to provide a marker of severity of autism symptoms relative to age and language level. The module a child can be given depends on his/her expressive language level, and thus will continue to be an important indicator of adaptive functioning for most children.

The dataset described here included children from various areas in the US, both urban and rural. Participants represented both consecutive clinic referrals and research participants. While this is likely a representative sample

for a North American clinical research center, it is worth examining how referral bias might have influenced these calibrated scores. Though the dataset was large ($N = 1,807$ assessments from children with ASD), its division into age/language cells for calibration resulted in a few small cell sizes. For example, children under age 5 who are not language delayed are unlikely to be referred for an evaluation unless they exhibit notable ASD symptomatology, so we would expect these cells to have a more limited distribution in the higher end of the range of ADOS scores. Another referral bias involved the tendency for children of higher severity to have more clinic reevaluations than those with less pronounced features of ASD. Indeed, the mean severity scores across the 18 calibration groups ranged from 6.64 (in young children with fluent speech) to 8.10 (in older children with phrase speech only), indicating that severity scores are still somewhat influenced by developmental level and referral bias.

After attempting a number of methods for standardizing ADOS scores, we believe that the present method of using ADOS diagnostic classifications to 'anchor' severity scores best controls for recruitment effects that would be present in any large clinical research sample, and therefore results in a metric more likely to be generalizable across datasets. If a cell in this calibration sample had predominantly high- or low-scoring children, this restricted range would only be assigned to severity scores associated with one classification (autism, ASD, or nonspectrum), allowing for more variability in other datasets across the other possible classifications. Ideally this method circumvents to some degree the inevitable effects of recruitment. Anchoring severity scores to ADOS classification instead of clinical diagnosis also avoids conflicting dimensional and diagnostic assignment. Within the present method, severity scores reflect ADOS raw totals regardless of the participant's diagnosis, so a child with a non-ASD best estimate diagnosis potentially could receive a score of 6 on the metric while a child with autism receives a 3, if the former child showed more autistic symptomatology relative to his/her age and language within that 45 min assessment than did the child with autism.

More work is needed to test the validity and utility of this calibrated severity metric. Module change, especially into Module 3 (fluent speech), may inflate an individual's severity score. Some longitudinal variation in these scores is expected, but the purpose of the metric is to measure change beyond typical variation in ASD. For this reason, the fact that approximately 20% of ASD assessments with 'autism' ADOS classifications receive the highest severity score of 10, creating a ceiling effect, was preferred over drawing out the distribution of the metric with the result of less meaningful differences between scores. We hope to

further examine patterns of severity score change over time in a longitudinal sample, identifying trajectory classes and the risk variables that predict class membership.

Another future direction is to calibrate the SA and RRB domains of the revised ADOS algorithms separately in order to measure severity within these symptom domains. This process will need to employ a different method of mapping raw scores onto a severity metric, due to the fact that each domain has a smaller range of possible raw totals than the overall score (with a maximum of only 8 points for the RRB domain).

Limitations

Although based on a large sample, this is not a metric of symptom severity in a “true” ASD population because ADOS data on such samples do not exist at present. As larger population studies become available, the metric should be recalibrated within those samples for a more accurate reflection of the distribution of ADOS scores in the ASD population.

These results also may be influenced by the historical period in which some of the data were collected. This sample grew over a 16-year period in which patterns in ASD identification evolved. As greater numbers of children are identified at earlier ages (thus including milder cases at younger ages), it is possible that severity scores might have been assigned differently to raw totals if only recently collected data were used.

Conclusion

The ADOS calibrated severity metric represents a step towards achieving greater comparability of scores across time, age, and module, and is less influenced by verbal IQ than raw scores. Therefore, it should provide a better measure of ASD severity than other methods currently available, including ADOS raw total scores. This metric must be replicated in a large independent sample. To test the validity of the metric, calibrated scores should be used to track observed changes in ASD severity against sources of convergent validity.

Calibrated scores could be used to predict outcome, changes in adaptive skills over time, and associations between severity of core features and clinical characteristics such as behavior problems, peer relationships, and school achievement. This metric may also prove useful in interpreting results from studies of the effectiveness of interventions, and in characterizing samples for genetic and neurobiological research. An important reminder, however, is that the calibrated severity metric is based on a relatively brief, office-based observation with a clinician, and thus is

only one part of a necessarily broader picture of the strengths and difficulties of a child with ASD.

Acknowledgments We gratefully acknowledge the help of Susan Risi, Kathryn Larson, Cristina Popa, and Mary Yonkovic, as well as the families that participated in this research. This study was funded by the National Institute of Mental Health (Validity of Diagnostic Measures for Autism Spectrum Disorders: NIMH RO1 MH066469) and an Autism Speaks Predoctoral Training Fellowship.

Disclosure C. Lord receives royalties for the ADOS; profits related to this study were donated to charity.

References

- Baird, G., Simonoff, E., Pickles, A., Chandler, S., Loucas, T., Meldrum, D., et al. (2006). Prevalence of disorders of the autism spectrum in a population cohort of children in South Thames—the Special Needs and Autism Project (SNAP). *Lancet*, *368*, 210–215. doi:10.1016/S0140-6736(06)69041-7.
- Bertrand, J., Mars, A., Boyle, C., Bove, F., Yeargin-Allsopp, M., & Decoufle, A. (2001). Prevalence of autism in a United States population: The Brick Township, New Jersey, investigation. *Pediatrics*, *108*(5), 1155–1161. doi:10.1542/peds.108.5.1155.
- Centers for Disease Control and Prevention. (2007). Prevalence of autism spectrum disorders—autism, developmental disabilities monitoring network, 14 sites, United States, 2002. *MMWR. Morbidity and Mortality Weekly Report*, *56*, 12–27.
- Chakrabarti, S., & Fombonne, E. (2005). Pervasive developmental disorders in preschool children: Confirmation of high prevalence. *The American Journal of Psychiatry*, *162*(6), 1133–1141. doi:10.1176/appi.ajp.162.6.1133.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale: Lawrence Erlbaum Associates.
- Constantino, J. N., Davis, S. A., Todd, R. D., Schindler, M. K., Gross, M. M., Brophy, S. L., et al. (2003). Validation of a brief quantitative measure of autistic traits: Comparison of the social responsiveness scale with the autism diagnostic interview-revised. *Journal of Autism and Developmental Disorders*, *33*(4), 427–433. doi:10.1023/A:1025014929212.
- de Bildt, A., Sytema, S., Ketelaars, C., Kraijer, D., Mulder, E., Volkmar, F., et al. (2004). Interrelationship between autism diagnostic observation schedule-generic (ADOS-G), autism diagnostic interview-revised (ADI-R), and the diagnostic and statistical manual of mental disorders (DSM-IV-TR) classification in children and adolescents with mental retardation. *Journal of Autism and Developmental Disorders*, *34*(2), 129–137. doi:10.1023/B:JADD.0000022604.22374.5f.
- DiLavore, P. C., Lord, C., & Rutter, M. (1995). The pre-linguistic autism diagnostic observation schedule. *Journal of Autism and Developmental Disorders*, *25*, 355–379. doi:10.1007/BF02179373.
- Elliot, C. D. (1990). *Differential abilities scale (DAS)*. San Antonio: Psychological Corporation.
- Gilliam, J. E. (1995). *Gilliam autism rating scale*. Austin: Pro-Ed.
- Gotham, K., Risi, S., Pickles, A., & Lord, C. (2007). The autism diagnostic observation schedule (ADOS): Revised algorithms for improved diagnostic validity. *Journal of Autism and Developmental Disorders*, *37*, 400–408. doi:10.1007/s10803-006-0280-1.
- Gotham, K., Risi, S., Dawson, G., Tager-Flusberg, H., Joseph, R., Carter, A., et al. (2008). A replication of the autism diagnostic observation schedule (ADOS) revised algorithms. *Journal of the American Academy of Child and Adolescent Psychiatry*, *47*(6), 643–651. doi:10.1097/CHI.0b013e31816bffb7.
- Joseph, R. M., Tager-Flusberg, H., & Lord, C. (2002). Cognitive profiles and social-communicative functioning in children with

- autism spectrum disorder. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 43(6), 807–821. doi:10.1111/1469-7610.00092.
- Krug, D. A., Arick, J. R., & Almond, P. J. (1980). Behavior checklist for identifying severely handicapped individuals with high levels of autistic behavior. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 21(3), 221–229. doi:10.1111/j.1469-7610.1980.tb01797.x.
- Lord, C., & Spence, S. (2006). Autism spectrum disorders: Phenotype and diagnosis. In S. Moldin & J. Rubenstein (Eds.), *Understanding autism: From basic neuroscience to treatment* (pp. 1–23). New York: Taylor and Francis.
- Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Jr, Leventhal, B. L., DiLavore, P. C., et al. (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, 205–223. doi:10.1023/A:1005592401947.
- Luyster, R., Gotham, K., Guthrie, W., Coffing, M., Petrak, R., Pierce, K., Bishop, S., Esler, A., Hus, V., Richler, J., Risi, S., & Lord, C. (2009). The autism diagnostic observation schedule—toddler module: A new module of a standardized diagnostic measure for ASD. *Journal of Autism and Developmental Disorders*, submitted.
- McCarthy, P. L., Cicchetti, D. V., Sznajderman, S. D., Forsyth, B. C., Baron, M. A., Fink, H. D., et al. (1991). Demographic, clinical and psychosocial predictors of the reliability of mothers' clinical judgments. *Pediatrics*, 88, 1041–1046.
- Mullen, E. (1995). *Mullen scales of early learning* (AGS Ed.). Circle Pines: American Guidance Service.
- Pickles, A., Starr, E., Kazak, S., Bolton, P., Papanikolaou, K., Bailey, A., et al. (2000). Variable expression of the autism broader phenotype: Findings from the extended pedigrees. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 41, 491–502. doi:10.1017/S0021963099005557.
- Rutter, M., Le Couteur, A., & Lord, C. (2003). *Autism diagnostic interview-revised-WPS* (WPS ed.). Los Angeles: Western Psychological Services.
- Schopler, E., Reichler, R. J., & Renner, B. R. (1986). *The Childhood Autism Rating Scale (CARS) for diagnostic screening and classification of autism*. Irvington: Irvington.
- South, M., Williams, B. J., McMahon, W. M., Owley, T., Filipek, P. A., Shernoff, E., et al. (2002). Utility of the Gilliam autism rating scale in research and clinical populations. *Journal of Autism and Developmental Disorders*, 32(6), 593–599. doi:10.1023/A:1021211232023.
- Sparrow, S., Balla, D., & Cicchetti, D. (1984). *Vineland adaptive behavior scales*. Circle Pines: American Guidance Service.
- Sparrow, S. S., Cicchetti, D. V., & Balla, D. A. (2005). *Vineland adaptive behavior scales* (2nd ed.). Circle Pines: American Guidance Service, Inc.
- Spiker, D., Lotspeich, L. J., Dimiceli, S., Myers, R. M., & Risch, N. (2002). Behavioral phenotypic variation in autism multiplex families: Evidence for a continuous severity gradient. *American Journal of Medical Genetics*, 114(2), 129–136. doi:10.1002/ajmg.10188.
- Szatmari, P., Bryson, S. E., Boyle, M. H., Streiner, D. L., & Duku, E. (2003). Predictors of outcome among high functioning children with autism and Asperger syndrome. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 44, 520–528. doi:10.1111/1469-7610.00141.
- Volkmar, F. R., Cicchetti, D. V., Dykens, E., Sparrow, S., Leckman, J. F., & Cohen, D. F. (1988). An evaluation of the autism behavior checklist. *Journal of Autism and Developmental Disorders*, 18, 81–97. doi:10.1007/BF02211820.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.