J Autism Dev Disord (2007) 37:613–627
DOI 10.1007/s10803-006-0280-1

ORIGINAL PAPER

The Autism Diagnostic Observation Schedule: Revised Algorithms for Improved Diagnostic Validity

Katherine Gotham · Susan Risi · Andrew Pickles · Catherine Lord

Published online: 16 December 2006

Abstract Autism Diagnostic Observation Schedule (ADOS) Modules 1–3 item and domain total distributions were reviewed for 1,630 assessments of children aged 14 months to 16 years with an autism spectrum disorder (ASD) or with heterogeneous non-spectrum disorders. Children were divided by language level and age to yield more homogeneous cells. Items were chosen that best differentiated between diagnoses and were arranged into domains on the basis of multifactor item-response analysis. Reflecting recent research, the revised algorithm now consists of two new domains, Social Affect and Restricted, Repetitive Behaviors (RRB), combined to one score to which thresholds are applied, resulting in generally improved predictive value.

Keywords Autism · Autism spectrum disorders · PDD-NOS · Diagnosis

Introduction

The Autism Diagnostic Observation Schedule (ADOS) is a semi-structured, standardized assessment of communication, social interaction, play, and imagination designed for use in diagnostic evaluations of individuals referred for a possible Autism Spectrum Disorder (ASD). For both research and clinical diagnostic purposes, the ADOS is intended to complement information obtained from developmental tests and a caregiver history, such as the Autism Diagnostic Interview-Revised (ADI-R; Rutter, LeCouteur, & Lord, 2003). The ADOS encompasses four modules, each with its own schedule of activities that allow examiners to observe behavior in participants of particular developmental and language levels, ranging from those with no expressive language to verbally fluent children and adults. Items are scored on a 4-point scale, with the highest scores of 2 and 3 collapsed in the algorithm in order to reduce impact of individual items.

To receive an ADOS classification of Autism or ASD, an individual's scores must meet separate cutoffs in a Communication domain, a Social domain, and a summation of the two. To date, the ADOS has been effective in categorizing children who definitely have autism or not, but has had lower specificity and sometimes sensitivity for distinctions involving children with milder ASDs (Lord et al., 2000; Bishop & Norbury, 2002; de Bildt et al., 2004). In the original norming sample for Modules 1–3, the ADOS generally achieved 94% correct classification. The exceptions were the ASD versus Non-spectrum Module 2 specificity of 87% and Module 3 sensitivity of 90%, and the Pervasive Developmental Disorder-Not Otherwise Specified (PDD-NOS) versus Non-spectrum Module 2 specificity of 88% and sensitivity of 89% and Module 3 sensitivity of 80% (Lord, Rutter, DiLavore, & Risi, 1999). That sample included 188 children and adolescents (recipients of Modules 1–3), with at most 21 participants in a

K. Gotham (✉) · S. Risi · 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
Division of Epidemiology and Health Science, University of Manchester, Manchester M13 9PT, United Kingdom

diagnostic group per module; the 2000 data were published with a request for replication in larger samples.

Despite the initial evidence for strong validity in classifying ASDs, several concerns can be raised about using the ADOS including floor and ceiling effects in the current algorithm totals and the effect of level of impairment. The ADOS scores in the norming sample had minimal association with verbal mental age (Lord et al., 1999), and divisions into modules ensured that two individuals functioning at the same mental age would receive the same schedule of activities and thus receive scores on the same items, regardless of their chronological age. However, an 8- and a 4-year-old who primarily used simple phrases would be scored similarly on the same ADOS algorithm, despite a clear difference in their levels of developmental impairment. In 2002, Joseph et al. (Joseph, Tager-Flusberg, & Lord, 2002) reported that ADOS social domain totals were correlated with level of cognitive impairment for preschool children and with the discrepancy between verbal and non-verbal IQ as measured on the Differential Ability Scales (DAS; Elliott, 1990). Data from de Bildt et al. (2004) suggested that, in a sample of children with mental retardation (MR), ADOS classifications appeared to be least valid for children with mild MR. It is unclear the exact role chronological age plays in the issue of impairment level; de Bildt et al. reported ADOS sensitivity increased with age in their mentally retarded sample (2004), while Lord et al. found the opposite effect in the original norming sample that had a smaller percentage of participants with MR (2000). Results from Bishop and Norbury (2002) suggest the ADOS may also be overinclusive with children with specific language impairments, though Noterdaeme et al. found excellent agreement between ADOS and clinicians' classification in their language-impaired sample (Noterdaeme, Mildenberger, Sitter, & Amorosa, 2002).

One approach to improve the sensitivity and specificity while possibly reducing the age and IQ effects of the ADOS is to divide the sample into smaller, more homogeneous cells by developmental level, language level, or age, and then create algorithms composed of the items that best differentiated between clinical diagnoses within each cell. Because our goal was to generate improvements that could be used with existing data, the current ADOS divisions by module were retained. An eventual goal in ASD identification is to integrate information from the ADOS and the ADI-R for individual cases; thus, at a minimum, we also wanted the new groupings proposed to be comparable to ADI-R distinctions of language level.

Previous factor analyses of the ADOS identified one factor underlying the social and communication domain items (Robertson, Tanguay, L'Ecuyer, Sims, & Waltrip, 1999; Lord et al., 1999, 2000), though separating the two domains yielded slightly higher specificity. Although considered separately in DSM-IV and ICD-10, several recent studies have suggested that non-verbal communication and social items often load onto the same factor (Robertson et al., 1999; Constantino et al., 2004). At issue, too, is the inclusion of restricted, repetitive behavior (RRB) items in an ADOS diagnostic total. Currently, these items appear on the algorithm but do not contribute to the total score that results in a spectrum/non-spectrum classification. This decision was based on the narrow window of time available to observe such behaviors in the context of the administration (Lord et al., 2000). However, recent findings suggest that RRB, even in the limited context of the ADOS, may make an independent contribution to diagnostic stability (Lord et al., 2006). Another reason for reviewing the current ADOS algorithms was to test whether including RRB items in the total score, so that they contribute to the total but are not required for a classification of autism, increased the validity of the measure.

The goal of the present research is to address these topics in order to improve the sensitivity and specificity of the ADOS algorithms for Modules 1 through 3, and to test the feasibility of employing items of similar conceptual content, though developmentally graded, in algorithms across all three modules used with children, allowing for easier comparison of cases. This endeavor is the first step of a larger project that aims to use ADOS algorithm scores from existing data to generate a calibrated metric of severity of autism, as independent as possible of current language levels. For ease in scoring, we wanted to create an algorithm with fewer items, chosen from items with the best possible diagnostic distinctions for increased predictive value of the measure and organized to remain as consistent as possible across developmental cells while maintaining or improving classification performance. Because we suspected floor and/or ceiling effects may occur in the current ADOS totals, we began with all available items as options for inclusion in a new algorithm, instead of attempting to adjust the item pool of the current algorithm. Our goal is to improve the usefulness of the ADOS in quantifying social-communicative deficits and in making more difficult diagnostic distinctions between ASD and other disorders.

Methods

Participants

Analyses were conducted on data from 1,139 different *participants*. Some participants had repeated assessments yielding a total of 1,630 *cases* (each case was defined by complete data from a contemporaneous ADOS, verbal IQ, and best estimate clinical diagnosis); thus one participant could provide data for two or three cases based on evaluations conducted at different points in time. From the sample of 1,630, 321 were given an ADOS precursor, the Pre-Linguistic ADOS (PL-ADOS), from which scores on identical items were recoded as Module 1 ADOS scores. The majority of participants completed a diagnostic evaluation at the University of Chicago Developmental Disorders Clinic or the University of Michigan Autism and Communication Disorders Center (UMACC). The rest participated in a longitudinal study conducted through TEACCH Centers at the University of North Carolina, Chapel Hill, and a clinic at the University of Chicago, or in recent, ongoing studies at UMACC, in which participants with non-ASD developmental delays, ASD-affected sibling pairs, or children between 12 and 36 months of age who fail a social-communication screener are recruited for a comprehensive evaluation. The sample was limited to participants aged 12 years or younger for Modules 1 and 2, and 16 years or younger for Module 3. The resulting age range of the sample is 14–192 months. Because older adolescents and adults with ASD were seen as a behaviorally distinct group that merited individual study, ADOS Module 4 recipients were excluded from the outset.

The final dataset included 912 *cases* with clinical diagnoses of autism (56% of entire sample), 439 with non-autism ASD (27%), and 279 with non-ASD developmental delays (17%). Within the non-spectrum sample of 279 *cases*, 115 had non-specific MR (41% of non-spectrum total), 58 were cases with language disorders (21%), 35 with oppositional defiant disorder, ADD and/or ADHD (12%), 38 with Down syndrome (14%), 16 with mood and/or anxiety disorders (6%), and 17 with an unspecified early delay (6%). Refer to Table 1 for a more detailed description of this sample.

Gender varied across module and diagnostic group from 57 to 86% male. Ethnicity across module and diagnostic group ranged from 71 to 91% Caucasian, 4–27% African American, 1–5% Asian American, 0–0.8% Native American, 0–2.2% biracial, and 0–0.6% other or unknown race.

Measures and Procedure

The ADOS was administered by a clinical psychologist or a trainee who had completed research training and met standard requirements for research reliability (Lord et al., 1999). A developmental hierarchy of psychometric measures, most frequently the Mullen Scales of Early Learning (MSEL; Mullen, 1995) and the DAS (Elliott, 1990) were used to determine IQ scores. Cognitive testing generally took place immediately before the ADOS administration. The ADI-R was available for 1,357 cases in our sample and the Vineland Adaptive Behavior Scales (VABS; Sparrow, Balla, & Cicchetti, 1984) for 1,409 cases. These two measures were administered together during a parent appointment that generally preceded the child assessment. Clinicians (usually a clinical psychologist and child psychiatrist) involved in each case together determined a best estimate diagnosis after review of all information. Clinic-referred participants received oral feedback and a written report without financial compensation. Participants who were recruited only for the purpose of research received compensation and a written summary of evaluation results. All procedures related to this research were approved by the Institutional Review Boards at the University of Chicago or the University of Michigan.

Inter-rater reliability on the ADOS was monitored through joint administration and scoring by two different examiners for at least 1 in 10 cases and, in some cases, through scoring of videotapes. Agreement remained at greater than 85%. Disagreements were resolved through discussion. Within this sample, 26 different examiners collected the data from the ADOS over 10 years.

Design and Analysis

The ADOS *domain means were compared* by module and diagnosis for the current sample and the original ADOS norming sample. *Domain total distributions* for this sample were generated within each module. When distributions appeared to exhibit floor or ceiling effects, items within that domain were evaluated to identify individual variables contributing to the effect. *Correlations* between ADOS totals and chronological age, verbal IQ, and verbal mental age were examined, and where possible, the *sample was divided* by age and language level within each module to yield cells with lower correlations between the ADOS totals and these variables. At that point, item distributions were considered within each cell in order to select those that best differentiated between diagnoses.

Table 1 Sample description

DX		Mod 1				Mod 2				Mod 3			
		N	Mean	*SD*	Range	*N*	Mean	*SD*	Range	*N*	Mean	*SD*	Range
Autism	age	592	54.64	26.83	14–144	188	79.60	29.27	28–143	132	100.81	29.35	42–183
	viq	592	31.28	18.52	2–103	188	60.82	21.73	22–127	132	85.29	23.18	31–159
	nviq	579	57.68	23.18	2–144	185	81.39	24.16	25–150	131	91.13	23.22	34–155
	vma	592	14.53	9.43	1–65	188	41.96	15.30	13–102	132	81.90	35.30	35–264
	nvma	589	28.24	14.20	2–110	186	62.75	25.54	17–155	131	91.47	31.93	25–165
	ADI social	526	20.92	5.82	1–30	137	20.54	6.02	2–30	93	19.33	5.64	1–29
	ADI comm-V	67	15.72	4.42	3–24	131	17.44	3.92	4–25	93	16.75	4.50	5–25
	ADI comm-NV	526	11.21	2.96	0–14	137	9.96	3.14	0–14	93	9.00	3.50	0–14
	ADI-RR	526	4.99	1.96	0–10	137	6.64	2.55	0–12	93	7.47	2.72	2–12
	ADOS social	592	11.28	2.22	0–14	188	10.48	2.39	5–14	132	8.83	2.71	0–14
	ADOS comm	592	6.31	1.63	0–10	188	7.59	1.64	2–10	132	4.85	1.80	0–8
PDD-NOS	age	160	40.88	17.01	15–107	91	60.65	21.57	28–130	188	100.29	30.02	45–172
	viq	160	51.40	22.07	7–108	91	72.40	17.78	28–121	188	99.47	21.50	49–151
	nviq	158	72.76	25.69	15–128	91	83.52	22.80	31–130	181	97.78	19.86	47–153
	vma	160	19.56	11.97	1–68	91	39.65	11.32	18–84	188	99.55	42.70	39–264
	nvma	159	28.76	14.87	7–78	91	50.11	17.81	23–134	181	97.27	33.10	37–190
	ADI social	148	14.11	6.46	0–28	69	12.06	6.42	2–29	133	15.11	7.49	0–29
	ADI comm-V	41	10.56	5.27	1–25	58	11.14	4.74	0–21	133	12.78	5.41	0–24
	ADI comm-NV	148	8.43	4.05	0–14	69	5.68	3.53	0–13	133	6.53	3.71	0–14
	ADI-RR	148	3.61	2.31	0–9	69	4.22	2.61	0–10	133	5.29	3.04	0–12
	ADOS social	160	7.71	3.50	0–14	91	6.62	3.37	0–14	188	6.02	2.88	0–14
	ADOS comm	160	3.97	2.09	0–9	91	5.26	1.95	1–10	188	3.15	1.68	0–8
Non-spectrum	age	135	43.79	22.09	14–129	61	69.85	30.71	37–143	83	103.75	29.98	51–192
	viq	135	57.99	24.89	10–113	61	74.84	21.92	24–120	83	90.70	21.90	41–139
	nviq	131	69.15	27.15	13–132	61	76.05	24.72	24–120	83	89.35	22.44	40–151
	vma	135	22.54	12.50	1–54	61	45.41	12.24	26–70	83	90.36	30.53	32–184
	nvma	134	27.63	12.85	4–76	61	50.20	14.94	19–93	83	91.86	31.38	34–129
	ADI social	123	9.06	6.59	0–26	51	10.35	6.25	0–28	77	9.58	6.91	0–24
	ADI comm-V	41	4.54	3.61	0–13	50	9.58	5.19	2–23	77	8.84	5.85	0–24
	ADI comm-NV	123	5.62	4.31	0–14	51	5.10	4.08	0–14	77	4.90	3.86	0–14
	ADI-RR	123	2.13	1.72	0–6	51	3.65	2.70	0–9	77	3.40	2.51	0–10
	ADOS social	135	3.93	3.64	0–14	61	2.44	2.21	0–9	83	3.29	2.44	0–9
	ADOS comm.	135	2.24	2.19	0–9	61	2.74	1.89	0–8	83	1.61	1.26	0–5

All ages are in months. *Age* Chronological Age, *viq* Verbal IQ, *nviq* Nonverbal IQ, *vma* Verbal Mental Age, *nvma* Nonverbal Mental Age, *ADI social* ADI Social Total, *ADI comm-V* ADI Communication Total for Verbal Ss, *ADI comm-NV* ADI Communication Total for Nonverbal Ss, *ADI-RR* ADI Restricted, Repetitive Behaviors Total, *ADOS social* ADOS Social Total, *ADOS comm* ADOS Communication Total

Each item in each cell was labeled as preferred or not preferred for inclusion in a new algorithm, with inclusion criteria generated from and applied to social-communication items, but not RRB items, which had an expected diversity (Bishop, Richler, & Lord, 2006). The criteria specified no more than 20% of autism cases scoring a zero on an item, and no more than 20% of non-spectrum cases scoring a 2 or 3. The former percentage was allowed to rise to 27–45% for two theoretically important items which performed well in some but not all of the cells (''Gestures'' for Module 3 and ''Shared Enjoyment'' for Modules 2 and 3). From this pool of preferred items, roughly equivalent items across modules were selected, so as to promote a conceptually uniform model across modules that would enhance inter-module comparisons.

Exploratory multi-*factor item response analysis* provided insight into the factor structure within each cell and was used to organize the items into new domains. All factor analyses reported here employed Mplus software (Muthen & Muthen, 1998) to address the ordinal nature of ADOS data. In an effort to balance getting the best fit by cell with having one model consistent across cells, factor loadings from promax oblique rotations were used to select better-performing items across developmental cells in a theoretically meaningful way. For example, where the item ''Pointing'' had failed to differentiate effectively between diagnoses in one cell, ''Response to Joint Attention'' replaced it; the item was theoretically similar (relating to shared interest) and loaded on the same factor. Goodness-of-fit was verified through

confirmatory factor analysis, and *logistic regression* used to examine the weighting of the two domains in view of the relative predictive value of scores from the different factors.

Domain total distributions of the new algorithm model were assessed for floor and ceiling effects, and *correlations* were generated between items and the remainder of the domain, as well as between items and participant characteristics like age and IQ. The ROC curves were calculated, and the *sensitivity and specificity* of the existing and newly revised ADOS algorithms compared within each cell. Since adjustment for the minority of subjects with multiple observations left the factor analysis results reported here largely unchanged, no adjustment has been made. Reported logistic regression coefficients for predicting diagnosis were adjusted using cluster robust standard errors, confidence intervals, and test statistics (Binder, 1983).

Results

Comparison of Domain Means

The ADOS domain total means and standard deviations were calculated for this sample in order to compare them to those of the original ADOS norming sample (Lord et al., 1999). For this comparison only, data from the original norming sample were removed from the current sample. As expected with a sample of the current size, mean differences in chronological age, verbal mental age, and non-verbal mental age between the module and diagnostic groups of the norming and current samples were statistically significant; however, they were clinically marginal. For example, the mean chronological age of Module 3 Autism groups was 8.45 years in the new sample ($N = 123$; $SD = 2.51$) and 9.14 years in the original norming sample ($N = 21$; $SD = 2.36$).

The ADOS domain means of the autism and ASD samples were similar, with a trend toward slightly lower means in Communication and Social domains for the Autism groups and slightly higher means in the Restricted-Repetitive domain for the non-autism ASD groups in the current larger sample. As examples, the mean combined social-communication totals in the Module 2 Autism groups were 18.38 in the norming sample ($N = 21$) and 18.11 in the current sample ($N = 171$), while the means for the Module 2 non-autism ASD groups were 11.83 ($N = 18$) in the norming sample and 11.94 ($N = 83$) in the current sample. On the whole, the similarities in domain distributions with greater numbers and a more diverse population suggest it would be appropriate to apply a new algorithm calibrated on this new sample to existing research databases.

Domain Total Distributions

In the original algorithm communication domain total, Module 1 scores of 8 were frequent (22.2% of Module 1 ASD sample), while scores of 9 or 10 were very rarely achieved (a total of 2.1% of Module 1 ASD cases received either score), implying an item set that provided little discrimination at the severe end of the spectrum. This effective range restriction was associated with the fact that 62% of the Module 1 ASD participants were non-verbal (i.e., participants used fewer than five words during the ADOS administration, as reflected in scores of 3 or 8 on Item 1: overall level of language [or a score of '4' on the equivalent PL-ADOS item]). For these children, only four items were scorable in the algorithm communication total. Scores of 9 and 10 were largely ineligible because the algorithm items ''Stereotyped /Idiosyncratic Use of Language'' and ''Frequency of Vocalization'' were unscorable for non-verbal participants, and thus did not contribute to the domain totals. Across all modules, distributions were broadened considerably when ''3'' codes were not recoded to ''2'' prior to algorithm calculation, but because standard use of the ADOS does not require reliable distinctions between these codes, our primary data analyses focused on continued use of this recoding. Because of the range restriction, we proposed the creation of distinct algorithms for verbal and non-verbal recipients of Module 1.

Correlations with Participant Characteristics

In Module 2, significant correlations between ASD participants' social-communication totals and their chronological age ($r = 0.24$, $p < 0.001$) and verbal IQ ($r = -0.34$, $p < 0.001$) occurred. Perusal of scatterplots provided evidence of curvilinear relationships between chronological age and ADOS social-communication totals, such that these variables were negatively related in children under age 5 (as age increased, ADOS scores decreased) and positively related in children over 5 (as age increased, ADOS scores increased). The children over 5 with phrase-speech-only seemed to represent a different group than children under 5 in Module 2, who may well acquire fluent speech as they get older. When Module 2 was split into ''Younger than 5'' and ''Greater or Equal to 5,'' the former group had no significant correlation between the social-communication total and age and verbal IQ (age: $r = -0.06$,

$p = 0.52$; VIQ: $r = -0.16$, $p = 0.08$); in the latter group, scores were still significantly correlated with these variables (age: $r = 0.16$, $p = 0.04$; VIQ: $r = -0.34$, $p < 0.001$), but less so in the case of age than before the division.

Division of the Modules

Dividing the sample by age in Module 3 did not produce more homogeneous samples. Dividing Module 3 recipients by language level was deemed not appropriate because past data had shown that ''Item 1: Overall Level of Language'' in Module 3 was particularly difficult to score reliably. Division based on specific item scores, such as ''Reporting of Events,'' met with little success. Thus, at this point, the cells for which revised algorithms have been formulated are Module 1, No Words; Module 1, Some Words; Module 2 Younger than 5; Module 2, 5 or Older; and Module 3 (Fig. 1).

Factor Analysis

Exploratory factor analysis was performed by cell (Fig. 1) with the 'preferred' items included from all domains. Item scores of 2 and 3 were collapsed and scores of 8 were labeled as missing data and excluded. Because ADOS data are ordinal and do not represent equal intervals, the analyses were run as ordinal probit item response models with Mplus Version 3.0 software. A Root Mean Square Error Approximation (RMSEA) of 0.08 or less is commonly taken as a satisfactory fit (Brown & Cudeck, 1993). The results shown in Table 2 indicate that 2-factor solutions generally fitted well, with items loading onto clear Social Affect (SA) and RRB factors that were positively correlated. Confirmatory factor analysis, that assigned each item to one of two factors, showed the 2-factor model to fit substantially better than the 1-factor model, with goodness-of-fit ratings ranging between Comparative Fit Index (CFI) of 0.94 (CFI between 0.9 and 1 indicating good fit; Skrondal and Rabe-Hesketh, 2004) and RMSEA of 0.08 in the Module 3 cell, to CFI of 0.97 and RMSEA of 0.09 in the Module 1, Some words cell.

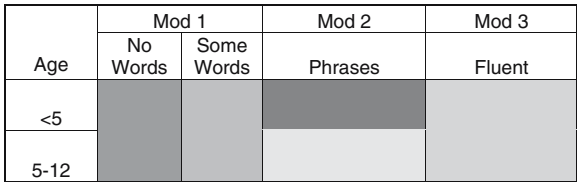

Fig. 1 Revised algorithm developmental cells

Thus, the final mapping of the new algorithm model includes a Social Affect domain and a Restricted-Repetitive domain (Table 2).

Although ''Stereotyped/Idiosyncratic Use of Words or Phrases'' was a communication domain item in the previous algorithms, it loaded onto the RRB factor and was thus included in that domain on the new algorithms.

The eigenvalues of a third factor, called ''Joint Attention,'' ranged across cells from 0.93 to 1.12 in exploratory factor analysis; this factor was comprised of pointing, gesturing, showing, initiating joint attention, and unusual eye contact in the Module 1, Some Words and both Module 2 groups, and response to joint attention, gesturing, showing, initiating joint attention, and unusual eye contact in the Module 1, No Words group. Statistics from confirmatory factor analysis were satisfactory (CFI ranged from 0.92 to 0.96; RMSEA from 0.06 to 0.09) in the four relevant developmental cells. The two-factor model, however, was more consistent across the five cells and more parsimonious. Although it was not included in the algorithm and overlaps with the SA factor, the Joint Attention factor was consistent across Modules 1 and 2 and therefore may be of interest to some researchers and clinicians.

Logistic Regression Check on Weighting Domains

Logistic regressions for autism versus not-autism (non-autism ASD and non-spectrum cases together), and ASD versus non-spectrum indicated that both the SA and RRB factors made significant independent contributions to the prediction of diagnosis. Since factor scores were not uniformly better at prediction of diagnosis than simple totals, we describe results for the simple item totals that would be ordinarily used in clinical practice. We report raw and standardized log-odds coefficients, the latter being easier to compare when the predictor variables have widely differing variability.

Item totals within SA and RRB factors were both predictive of diagnosis. For children with autism versus all other groups, the raw partial log-odds coefficients were 0.29 (C.I. 0.26, 0.32; $z = 16.53$; standardized coefficient = 1.74) for SA and 0.36 (C.I. 0.29, 0.44; $z = 9.30$; standardized coefficient = 0.84). For Autism and PDD versus no PDD, the raw partial log-odds coefficients were 0.25 (C.I. 0.20, 0.29; $z = 10.52$; standardized coefficient = 1.47) for SA and 0.51 (C.I. 0.38, 0.64; $z = 7.92$; standardized coefficient = 1.20).

While both factors were predictive for both comparisons, the standardized coefficients and z-scores are

Table 2 Revised algorithm mapping

Domains		Mod 1, No Words $N = 495$	Factor Loadings	Mod 1, Some Words $N = 351$	Factor Loadings	Mod 2 Younger $N = 137$	Factor Loadings	Mod 2 Older $N = 192$	Factor Loadings	Mod 3 $N = 398$	Factor Loadings
Social affect		Unusual Eye Contact	0.64	Unusual Eye Contact	0.53	Unusual Eye Contact	0.55	Unusual Eye Contact	0.76	Unusual Eye Contact	0.51
		Gaze and Other Behaviors	0.76	Gaze and Other Behaviors	0.47	Amount of Social Communication	0.80	Amount of Social Communication	0.89	Amount of Social Communication	0.81
		Facial Expressions	0.85	Facial Expressions	0.68	Facial Expressions	0.70	Facial Expressions	0.82	Facial Expressions	0.67
		Frequency of Vocalization	0.69	Frequency of Vocalization	0.84	Quality of Rapport	0.80	Quality of Rapport	0.55	Quality of Rapport	0.72
		Shared Enjoyment	0.71	Shared Enjoyment	0.62	Shared Enjoyment	0.89	Shared Enjoyment	0.79	Shared Enjoyment	0.84
		Quality of Social Overtures	0.74	Quality of Social Overtures	0.65	Quality of Social Overtures	0.90	Quality of Social Overtures	0.68	Quality of Social Overtures	0.65
		Response to Joint Attention	0.60	Pointing	0.82	Pointing	0.66	Pointing	0.68	Conversation	0.73
		Gestures	0.73	Gestures	0.72	Gestures	0.34	Gestures	0.71	Gestures	0.71
		Showing	0.69	Showing	0.79	Showing	0.64	Showing	0.72	Quality of Social Response	0.60
		Initiation of Joint Attention	0.77	Initiation of Joint Attention	0.73	Initiation of Joint Attention	0.72	Initiation of Joint Attention	0.73	Reporting of Events	0.65
	Eigen value	7.1		8.3		6.8		7.8		6.2	
Restricted repetitive behaviors		Intonation	0.44	Stereotyped Language	0.70	Stereotyped Language	0.56	Stereotyped Language	0.55	Stereotyped Language	0.60
		Unusual Sensory Interest	0.78	Unusual Sensory Interest	0.70	Unusual Sensory Interest	0.45	Unusual Sensory Interest	0.61	Unusual Sensory Interest	0.44
		Repetitive Interests	0.44	Repetitive Interests	0.69	Repetitive Interests	0.43	Repetitive Interests	0.84	Highly Specific Topics	0.63
		Hand Mannerisms	0.66	Hand Mannerisms	0.63	Hand Mannerisms	0.99	Hand Mannerisms	0.37	Hand Mannerisms	0.48
	Eigen value	1.5		1.3		1.5		1.3		1.7	
	RMSEA	0.05		0.04		0.07		0.06		0.06	
	Rho	0.49		0.59		0.56		0.58		0.38	

Item have been abbreviated from the Western Psychological Services ADOS item names. Refer to the key from Fig.6 in the ADOS Manual (Lord et al., 1999) for complete names. *RMSEA* Root Mean-Square Error Approximation (values 0.08 or less indicate a good fit). *Rho* correlation between Social Affect and Restricted Repetitive Behaviors factors

Items from the 2000 algorithm not included in new algorithm: Stereotyped/Idiosyncratic Use of Language, Use of Other's Body to Communicate, and Pointing (Mod 1 No Words), Use of Other's Body to Communicate, Response to Joint Attention (Mod 1 Some Words); Amount of Social Overtures/Maintenance of Attention, Conversation, and Quality of Social Response (Mod 2 Younger and Older); Insight, and Compulsions and Rituals (Mod 3)

lower for RRB than SA. In addition, it is interesting to note that, for the SA factor, the log-odds coefficients are similar for the different factors, but for the RRB factor, the coefficient for the autism and PDD versus no-PDD comparison appears to be larger than that for the autism versus all other groups comparison.

Item Correlations with Domain Totals, Chronological Age, Mental Age, and IQ

Item-'rest' correlations (domain scores minus the particular item) were significant for each algorithm item in each developmental cell; they ranged from 0.45 to 0.78 in the SA domain and 0.27–0.53 in the RRB domain. The two domains were significantly correlated with each other (0.34–0.57 by cell). Internal consistency was assessed using Cronbach's alpha (Cronbach, 1951). Cronbach's alphas were consistently highest for the SA domain (0.87–0.92 by developmental cell) and ranged from 0.51 to 0.66 in the Restricted, Repetitive domain.

Item correlations with age and verbal mental age were also reviewed. ''Intonation'' in Module 1, No Words was the only item correlated above 0.30 with chronological age ($r = 0.45$, $p < 0.001$). Seven items across cells showed correlations with verbal mental age greater than 0.30; five of these items applied only to the Module 1, Some Words cell (ranging from ''Unusual Eye Contact,'' $r = -0.32$, $p < 0.001$ to ''Showing,'' $r = -0.39$, $p < 0.001$). Clearly, the delineation of children with ''Some Words'' in Module 1 still yields a heterogeneous group, in which social skills are related to children's language abilities, ranging from a few single words to the use of occasional phrases.

Sensitivity and Specificity

Receiver Operating Characteristic (ROC) curves (Siegel, Vukicevic, Elliott, & Kraemer, 1989) were run to obtain the sensitivity and specificity of both the old and the new algorithms by cell. For the new algorithms, ROC curves were run twice, for items from the SA factor alone and then for the sum of items from the SA and RRB factors. As in the past, scores of 3 were recoded to 2 for this procedure. Cases with acceptable missing data (for example, an '8' on the 'stereotyped speech' item in Module 1, Some Words) were included (contributing a zero score), but 56 cases were excluded because of other missing data from items comprising either the old or new algorithm, for a resulting N of 1,574. The new algorithm includes the item ''Integration of Gaze and Other Behaviors during Social Overtures'' in Module 1; because the children who received the PL-ADOS did not have this item available and yet represented a clinically important group, scores on the item ''Unusual Eye Contact'' were substituted for the missing item data in PL-ADOS recipients. The inclusion of the PL-ADOS cases greatly reduced specificity in the Module 1, No Words cell, the most obvious reason being the inclusion of children with very low non-verbal mental age in the Early Diagnosis sample (Lord et al., 2006), all of whom initially were assessed using the PL-ADOS. Because evaluating children with low non-verbal mental age is a reality in clinical practice, sensitivity and specificity were generated for all of Module 1, No Words cases, but were reported separately for those with non-verbal mental ages of 15 months or lower and those with non-verbal mental ages above 15 months for comparison (Table 3). Another point to note in Table 3 is that ''ASD'' is often reported in literature as including the Autism and non-autism ASD cases, whereas we have provided separate comparisons in this table of Autism versus Non-spectrum cases, and non-autism ASD cases (PDD-NOS and Asperger Disorder) versus Non-spectrum cases. This was done to give a true indication of how well the measure performs within the most conservative diagnostic groupings.

For Autism versus Non-spectrum, and for ASD versus NS (Table 3), the new and old algorithms perform approximately equally well in terms of sensitivity, with the new algorithm showing slightly reduced sensitivity in some cells and notable gains in others (Module 1, Some words; AUT versus NS and ASD versus NS). For non-autism ASD versus Non-spectrum, sensitivity of the new algorithm is somewhat lower in Module 1, No Words (as was necessary to raise specificity), but shows improvement from the old algorithm in the higher-functioning Modules 1 (AUT versus NS) and 2 (ASD versus NS) cells.

The new algorithm shows substantial gains in specificity in each of the diagnostic categories. Module 1, No Words (both non-verbal mental age groups) improve in each diagnostic comparison; the specificity of both Module 2 groups improves for non-autism ASD versus NS.

Overall, the first factor by itself tends to perform somewhat less well, so a summation of both domain totals are recommended to complete a total algorithm score. Analyses also were rerun including scores of 3 to see if using a broader distribution resulted in greater predictive value; there was little impact on sensitivity and specificity in comparison with the new totals with recoded 3's. Further information about cut-offs using ''3's'' is available from the authors.

Table 3 Sensitivities and specificities of current and revised algorithms

N = 1157	Meets Comm-Soc for Aut		New SA + RRB		New SA only	
	Sens	Spec	Sens	Spec	Sens	Spec
AUT versus NS						
Mod 1, no words, nvma ≤ 15 AUT = 69 NS = 16	100	19	97	50 (16)	96	50 (14)
Mod 1, no words, nvma > 15 AUT = 306 NS = 33	97	91	95	94 (16)	89	94 (14)
Mod 1, some words, AUT = 201 NS = 76	88	96	97	91 (12)	91	93 (11)
Mod 2 younger, AUT = 58 NS = 30	97	93	98	93 (10)	95	97 (9)
Mod 2 older, AUT = 126 NS = 30	96	97	98	90 (9)	92	97 (9)
Mod 3 AUT = 129 NS = 83	86	89	91	84 (9)	85	87 (8)
N = 685	Meets Comm-Soc for ASD		New SA + RRB		New SA only	
	Sens	Spec	Sens	Spec	Sens	Spec
Non-Autism ASD versus NS						
Mod 1, no words, nvma ≤ 15 PDD-NOS = 20 NS = 16	95	6	95	19 (11)	90	12 (9)
Mod 1, no words, nvma > 15 PDD-NOS = 51 NS = 33	88	67	82	79 (11)	80	76 (9)
Mod 1, some words, PDD-NOS = 75 NS = 76	67	84	77	82 (8)	75	79 (6)
Mod 2 younger, PDD-NOS = 49 NS = 30	76	70	84	77 (7)	80	63 (5)
Mod 2 older, PDD-NOS = 36 NS = 30	86	77	83	83 (8)	72	77 (6)
Mod 3 PDD-NOS = 186 NS = 83	68	77	72	76 (7)	61	78 (6)

Numbers in parentheses indicate best cut-off identified in ROC curves

Comm-Soc Communication+Social Cut-offs from 2000 norms, *SA* New (2006) Social Affect Domain, *RRB* New (2006) Restricted Repetitive Behaviors, *nvma* nonverbal mental age in months. Non-Autism ASD includes PDD-NOS and 3 cases of Asperger's Disorder

Discussion

With a much larger, more diverse sample (in terms of participants and examiners), both domain means and sensitivity and specificity remained similar to the original norming data, indicating that the ADOS continues to be a valid and reliable measure. Used with the original norming sample, any new algorithm was unlikely to improve on the old, since the latter had been chosen to best classify that particular sample. The new sample, being so much larger, offers less scope for overfitting, and thus achieving artifactual high levels of classification success. Nonetheless, as intended, the algorithm changes described here increase specificity in classifying non-autism ASD in lower functioning populations, evidenced by the 12–31% increase in specificity for children without any words (depending on non-verbal mental age) and the modest gain in specificity for older children who have not progressed beyond phrase speech.

A more homogeneous algorithm has been achieved, with similar items used across developmental cells to allow for easier comparison of ADOS scores within and between individuals. This is a step closer to the use of the ADOS as a measure of severity. The inclusion of repetitive behavior items in an algorithm model that is relatively uniform across developmental cells will be useful in the derivation and application of the future severity metric.

With the use of the proposed model, all items appearing on the algorithm will contribute to a single score with two classification thresholds, one for Autism and one for ASD. The existing social and communication domains were merged, as proposed in previous research (Lord et al., 1999, 2000; Robertson et al., 1999), and only very strong, salient factors were retained. Because longitudinal data suggest that trajectories between social communication and RRB profiles are different (Lord et al., 2006), the two-factor solution chosen from these analyses adds to the clinical utility of this diagnostic instrument. Inclusion of the RRB domain did not improve predictive value of the ADOS in differentiating individuals with autism from those with PDD-NOS, though surprisingly it aided in distinguishing PDD-NOS from non-spectrum cases.

Clearly some of the many goals for this algorithm revision were more difficult to achieve: the specificity of classification in children with non-verbal ages 15 months and younger remained weak. For these children, ADOS cut-offs do not reliably differentiate Autism or ASD from other disorders. It may be that expectations of interaction in the ADOS are too high for passive, low-functioning children (Hepburn, Lord, John, & Rogers, Submitted). A version of the ADOS employing novel

tasks for infants and toddlers is now being piloted at the UMACC to address the diagnostic needs of very young and/or more severely delayed toddlers.

Although predictive value of the ADOS for children with autism was strong across all groups, sensitivity for non-autism ASD across modules was lower than desired even in children with non-verbal ages above 15 months. Correlations between the revised totals and age, IQ, mental age, ADI-R, and former ADOS totals are reported here (see Tables 4, 5, 6). Division of the sample into cells accomplished the goal of creating algorithms independent of age effects (except for Module 1), and minimizing the effect of verbal IQ across modules. Greater association between ADOS and cognitive scores remained in the Module 1 cells relative to other developmental cells; this was expected due to the role of social communication in the measurement of cognitive skill in very young or low-functioning children. In fact, some of the earliest MSEL items in the language domain, as well as in other developmental tests (Bayley, 1993), overlap with social-communication items on the ADOS (e.g., MSEL item 'Responds to voice and face by vocalizing'). Severity of expressive language impairment (though not so much current language functioning) continues to influence our interpretation of autistic symptomatology, and even with Module 1 divided further on language level, the relationship between ADOS domain scores and verbal IQ remained.

The new algorithm did not greatly improve the distinction between autism and other ASDs in the ADOS. It continues to be the case that social-communication deficits within ASD (Constantino et al., 2004; Gilmour, Hill, Place, & Skuse, 2004) and shared by ASD and other developmental or psychiatric disorders (Bishop & Norbury, 2002) appear to represent a continuous dimension. Our next step is to generate such scores based on calibrations across modules. Because we found virtually the same total distribution and predictive value when including scores of 3 in the new algorithm, the calibration effort likely will recode these to 2's as is common practice. For those who want to increase variability in their ADOS data and who have become reliable on coding 3's as well as 2's (Rogers, Hepburn & Wehner, 2004), 3's remain an option to increase variation, particularly in treatment studies that look for changes in individuals.

Limitations

Because different thresholds were necessary to attain the best sensitivity and specificity within each developmental cell, calibration is necessary to achieve the final goal of providing a simple way to compare cases across modules. Although the goal of creating algorithms more similar across modules was met, it is ultimately limited by the fact that children receiving a Module 2, for example, still complete different tasks than do those in Module 3. Shifts in modules clearly add complexity in interpreting data, but there is little alternative to the fact that adequate social behavior differs with chronological age, language level, and examiner expectations, and thus must be measured by different modules of tasks. To maintain the validity of this measure, it is important that the appropriate module is selected (see also Klein-Tasman, Risi, & Lord, in press).

The sensitivity and specificity of the instrument may vary in different clinics and research centers due to the skill of the examiner, sequence of administration, and other factors; therefore, we might not see the same predictive value of the instrument in other clinical or research databases (see also Risi et al., 2006). A further limitation of this study is the relatively small numbers of non-spectrum participants by module.

The research protocol described here was unable to divide recipients of Module 3 into more homogeneous groups based on chronological age or language level. Comparison of item distributions between younger and older recipients of Module 3 revealed few age differences in children age 5 and over; children under 5 who received Module 3 exhibited some mean differences on an item level, but this group represents such a small minority of Module 3 recipients that we felt it unnecessary to divide the sample on that basis alone. The sensitivity and specificity of both old and new ADOS classifications are generally lower in Module 3 than in other developmental cells. Ideally, study of higher-functioning children and verbal adolescents will lead to better understanding of autism and ASDs in these populations and perhaps inform decisions on new tasks and/or scored items for future revision of this diagnostic schedule. Modifications of the ADOS for older children and adults with single words or phrase speech are also needed in order to present more age-appropriate tasks and materials to lower-functioning individuals while preserving standardization. The extent to which one can measure more subtle social-cognitive differences in one-to-one interaction with an adult in an office visit is unknown. Information from measures such as the VABS, ADI-R, Social Responsiveness Scale (SRS; Constantino et al., 2003), and Pervasive Developmental Disorder Behavior Inventory (PDDBI; Cohen, Schmidt-Lackner, Romanczyk, & Sudhalter, 2003), that allow consideration of the information from a broader range of contexts, may be critical.

Table 4 Correlations between revised algorithm totals and participant characteristics

			VIQ	NVIQ	VMA	NVMA	AGE
Mod 1 no words nvma ≤ 15	SA total	*r*	–0.49	–0.24	–0.43	–0.01	0.18
		p	0.00	0.02	000	0.93	0.06
		N	104	99	105	105	105
	RRB total	*r*	–0.40	–0.28	–0.06	0.02	0.28
		p	0.00	0.01	0.55	0.87	0.00
		N	104	99	105	105	105
	SA + RRB	*r*	–0.551	–0.305	–0.337	0.002	0.270
		p	0.000	0.002	0.000	0.981	0.005
		N	104	99	105	105	105
Mod 1 no words nmva > 15	SA total	*r*	–0.50	–0.23	–0.38	–0.02	0.14
		p	0.00	0.00	0.00	0.71	0.00
		N	390	377	390	390	390
	RRB total	*r*	–0.44	–0.46	–0.17	–0.00	0.43
		p	0.00	0.00	0.00	0.94	0.00
		N	390	377	390	390	390
	SA + RRB	*r*	–0.56	–0.36	–0.36	–0.01	0.25
		p	0.00	0.00	0.00	0.00	0.00
		N	390	377	390	390	390
Mod 1 words	SA total	*r*	–0.55	–0.25	–0.39	–0.04	24
		p	0.00	0.00	0.00	0.47	0.00
		N	352	346	352	350	352
	RRB total	*r*	–0.42	–0.34	–0.24	–0.06	0.27
		p	0.00	0.00	0.00	0.25	0.00
		N	352	346	352	350	352
	SA + RRB	*r*	–0.57	–0.31	–0.39	–0.05	0.28
		p	0.00	0.00	0.00	0.35	0.00
		N	352	346	352	350	352
Mod 2 younger	SA total	*r*	–0.24	–0.05	–0.23	0.00	0.06
		p	0.00	0.57	0.01	0.99	0.51
		N	137	130	137	136	137
	RRB total	*r*	–0.18	–0.05	–0.19	–0.04	–0.01
		p	0.04	0.57	0.03	0.62	0.90
		N	137	130	137	136	137
	SA + RRB	*r*	–0.25	–0.06	–0.24	–0.02	0.04
		p	0.00	0.52	0.00	0.87	0.63
		N	137	130	137	136	137
Mod 2 older	SA total	*r*	–0.25	0.24	–0.22	0.21	0.06
		p	0.00	0.00	0.00	0.00	0.39
		N	192	187	192	191	192
	RRB total	*r*	–0.29	–0.05	–0.18	0.07	0.11
		p	0.00	0.54	0.02	0.38	0.13
		N	192	187	192	191	192
	SA + RRB	*r*	–0.29	0.17	–0.23	0.19	0.08
		p	0.00	0.02	0.00	0.01	0.25
		N	192	187	192	191	192
Mod 3	SA total	*r*	–0.33	–0.18	–0.27	–0.15	–0.02
		p	0.00	0.00	0.00	0.00	0.71
		N	398	385	398	392	398
	RRB total	*r*	–0.07	–0.12	–0.04	–0.09	–0.03
		p	0.18	0.02	0.38	0.09	0.51
		N	398	385	398	392	398
	SA + RRB	*r*	–0.30	–0.19	–0.24	–0.15	–0.03
		p	0.00	0.00	0.00	0.00	0.60
		N	398	385	398	392	398

VIQ verbal IQ, *NVIQ* nonverbal IQ, *VMA* verbal mental age, *NVMA* nonverbal mental age, *AGE* chronological age, *SA total* Social Affect total, *RRB total* Restricted, Repetitive Behaviors total, *SA + RRB* Combined new algorithm total

A comparison of ADOS classification to clinical diagnosis, which was done to generate the sensitivity and specificity numbers reported here, is confounded by the fact that the two classifications are not independent, as the ADOS was one of the tools used to make the clinical best estimate diagnosis. When constructing an entirely new algorithm for a new instrument, an entirely independent validation criterion is desirable as proof of validity. However, when revising an algorithm, the concern is to identify improved performance over an existing algorithm, each being measured against the best available

Table 5 Correlations between revised algorithm totals and ADI-R domain totals

			ADI-R	Soc	CommV	CommNV	RR	Onset
Mod 1 no words nvma ≤ 15	SA total	*r*		0.40		0.11	0.26	0.05
		p		0.00		0.26	0.01	0.648
		N	100		0	100	100	100
	RRB total	*r*		0.42		0.03	0.31	0.17
		p		0.00		0.75	0.00	0.09
		N	100		0	100	100	100
	SA + RRB	*r*		0.50		0.10	0.34	0.12
		p		0.00		0.33	0.00	0.23
		N	100		0	100	100	100
Mod 1 no words nmva > 15	SA total	*r*		0.49	-0.02	0.40	0.24	0.17
		p		0.00	0.97	0.00	0.00	0.00
		N	356		6	358	356	356
	RRB total	*r*		0.43	0.20	0.27	0.44	0.38
		p		0.00	0.71	0.00	0.00	0.00
		N	356		6	356	356	356
	SA + RRB	*r*		0.55	0.08	0.42	0.36	0.29
		p		0.00	0.89	0.00	0.00	0.00
		N	356		6	356	356	356
Mod 1 words	SA total	*r*		0.60	0.65	0.60	0.42	0.17
		p		0.00	0.00	0.00	0.00	0.00
		N	307		140	307	307	307
	RRB total	*r*		0.48	0.65	0.48	0.50	0.20
		p		0.00	0.00	0.00	0.00	0.00
		N	307		140	307	307	307
	SA + RRB	*r*		0.63	0.71	0.62	0.49	0.19
		p		0.00	0.00	0.00	0.00	0.00
		N	307		140	307	307	307
Mod 2 younger	SA total	*r*		0.61	0.64	0.61	0.25	0.07
		p		0.00	0.00	0.00	0.01	0.45
		N	113		102	113	113	113
	RRB total	*r*		0.33	0.42	0.29	0.44	-0.01
		p		0.00	0.00	0.00	0.00	0.91
		N	113		102	113	113	113
	SA + RRB	*r*		0.60	0.64	0.58	0.34	0.05
		p		0.00	0.00	0.00	0.00	0.58
		N	113		102	113	113	113
Mod 2 older	SA total	*r*		0.51	0.50	0.45	0.30	0.22
		p		0.00	0.00	0.00	0.00	0.01
		N	137		132	137	137	137
	RRB total	*r*		0.39	0.36	0.29	0.46	0.21
		p		0.00	0.00	0.00	0.00	0.01
		N	137		132	137	137	137
	SA + RRB	*r*		0.53	0.52	0.45	0.39	0.25
		p		0.00	0.00	0.00	0.00	0.00
		N	137		132	137	137	137
Mod 3	SA total	*r*		0.41	0.41	0.38	0.29	0.19
		p		0.00	0.00	0.00	0.00	0.00
		N	299		299	299	299	299
	RRB total	*r*		0.19	0.31	0.24	0.39	0.08
		P		0.00	0.00	0.00	0.00	0.19
		N	299		299	299	299	299
	SA + RRB	*r*		0.40	0.45	0.39	0.37	0.18
		p		0.00	0.00	.00	0.00	0.00
		N	299		299	299	299	299

Soc ADI-R Reciprocal Social Interaction Total, *CommV* ADI-R Communication (Verbal) Total, *CommNV* ADI-R Communication (Nonverbal) Total, *RR* ADI-R Restricted, Repetitive Behaviors Total, *Onset* ADI-R Abnormality of Development Before 36 months Total, *SA total* Social Affect total, *RRB total* Restricted, Repetitive Behaviors total, *SA + RRB* Combined new algorithm total

criterion. Lord et al. (2006) have shown in a longitudinal study that clinical judgment, the ADI-R, and the ADOS all made independent contributions in predicting long-term best-estimate diagnoses. No single source can be considered as either a gold standard or the best possible criterion. This would suggest that to calibrate the algorithm against a criterion diagnosis that excluded information from the ADOS would be to calibrate it against a potentially inferior criterion. The fact that the best-estimate diagnosis was not independent of the ADOS might potentially upwardly bias the absolute performance of an ADOS algorithm;

Table 6 Correlations between revised algorithm totals and previous ADOS algorithm totals

		ADOS	Soc	Comm	Soc-co	Play	RR
Mod 1 no words nvma ≤ 15	SA total	*r*	0.95	0.52	0.94	0.18	0.37
		p	0.00	0.00	0.00	0.07	0.00
		N	105	105	105	105	105
	RRB total	*r*	0.32	0.26	0.35	0.22	0.89
		p	0.00	0.01	0.00	0.03	0.00
		N	105	105	105	105	105
	SA + RRB	*r*	0.83	0.50	0.84	0.24	0.70
		p	0.00	0.00	0.00	0.02	0.00
		N	105	105	105	105	105
Mod 1 no words Nmva > 15	SA total	*r*	0.97	0.80	0.97	0.55	0.44
		p	0.00	0.00	0.00	0.00	0.00
		N	90	390	390	390	390
	RRB total	*r*	0.41	0.38	0.43	0.47	0.90
		p	0.00	0.00	0.00	0.00	0.00
		N	390	390	390	390	390
	SA + RRB	*r*	0.91	0.76	0.92	0.61	0.69
		p	0.00	0.00	0.00	0.00	0.00
		N	390	390	390	390	390
Mod 1 words	SA total	*r*	0.96	0.87	0.98	0.59	0.52
		p	0.00	0.00	0.00	0.00	0.00
		N	352	352	352	352	352
	RRB total	*r*	0.52	0.61	0.59	0.50	0.92
		p	0.00	0.00	0.00	0.00	0.00
		N	352	352	352	352	352
	SA + RRB	*r*	0.93	0.88	0.96	0.63	0.70
		p	0.00	0.00	0.00	0.00	0.00
		N	352	352	352	352	352
Mod 2 younger	SA total	*r*	0.96	0.85	0.97	0.52	0.45
		p	0.00	0.00	0.00	0.00	0.00
		N	137	137	137	137	137
	RRB total	*r*	0.53	0.61	0.59	0.43	0.95
		p	0.00	0.00	0.00	0.00	0.00
		N	137	137	137	137	137
	SA + RRB	*r*	0.94	0.87	0.96	0.56	0.67
		p	0.00	0.00	0.00	0.00	0.00
		N	137	137	137	137	137
Mod 2 older	SA total	*r*	0.96	0.89	0.98	0.46	0.47
		p	0.00	0.00	0.00	0.00	0.00
		N	192	192	192	192	192
	RRB total	*r*	0.59	0.61	0.63	0.41	0.95
		p	0.00	0.00	0.00	0.00	0.00
		N	192	192	192	192	192
	New total	*r*	0.94	0.90	0.97	0.49	0.67
		p	0.00	0.00	0.00	0.00	0.00
		N	192	192	192	192	192
Mod 3	SA total	*r*	0.95	0.84	0.98	0.47	0.27
		p	0.00	0.00	0.00	0.00	0.00
		N	398	398	398	398	398
	RRB total	*r*	0.39	0.44	0.43	0.12	0.91
		p	0.00	0.00	0.00	0.02	0.00
		N	398	398	398	398	398
	New total	*r*	0.93	0.85	0.96	0.43	0.52
		p	0.00	0.00	0.00	0.00	0.00
		N	398	398	398	398	398

Soc ADOS Social Total (2000), *Comm* ADOS Communication Total (2000), *Soc-Co* ADOS Social-Communication Combined Total (2000), *Play* ADOS Play Total (2000), *RR* ADOS Restricted, Repetitive Behaviors Total (2000), *SA total* Social Affect total, *RRB total* Restricted, Repetitive Behaviors total, *SA + RRB* Combined new algorithm total

however, its influence on the relative performance would be slight. By contrast the much larger sample size of this study makes it less prone to the upward bias in absolute performance of previous studies that can arise from over-fitting.

Conclusions

The satisfactory performance of the revised algorithm found here must be replicated in other research samples before it replaces the existing ADOS algorithm. It

can be calculated currently by adding scores from the items listed under the relevant developmental cell in Table 2, and applying the parenthetical thresholds for Autism or ASD from Table 3. Pending replication and the future calibration project, we expect a new published version of the algorithm to be provided by Western Psychological Services.

The ADOS has begun to be used in relation to neurobiological measures (Critchley et al., 2000; Schultz et al., 2000; Klin, Jones, Schultz, Volkmar & Cohen, 2002) and continues to contribute to improved diagnosis in conjunction with the ADI-R (Risi et al., 2006). Nevertheless, researchers and clinicians must bear in mind that this measure is not a replacement for a historical account by a caregiver or for the diagnosis of a well trained, experienced clinician. Replication across sites and across other well defined populations with and without ASD and further explorations into how we can best organize time-limited, clinician-structured observations of social-communication behavior to better understand and treat ASD are all much needed.

Acknowledgments We gratefully acknowledge the help of Bennett Leventhal, Edwin Cook, Christina Corsello, Amy Esler, Somer Bishop, Danielle Guerin, Kathryn Larson, Jackie Preston Opatik, and Mary Yonkovit, 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). Some authors (C.L., S.R.) receive royalties for the ADOS; profits related to this study were donated to charity.

References

Bayley, N. (1993). *Bayley scales of infant development* (2nd ed.). San Antonio: The Psychological Corporation.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, *51*, 279–292.

Bishop, D. V., & Norbury, C. F. (2002). Exploring the borderlands of autistic disorder and specific language impairment: a study using standardized diagnostic instruments. *Journal of Child Psychology and Psychiatry & Allied Disciplines*, *43*(7), 917–929.

Bishop, S. L., Richler, J., & Lord, C. (2006). Restricted and repetitive behaviors and non-verbal IQ in children with autism spectrum disorders. *Journal of Child Neuropsychology, 12*, 247–267.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Cohen, I. L., Schmidt-Lackner, S., Romanczyk, R., & Sudhalter, V. (2003). The PDD behavior inventory: A rating scale for assessing response to intervention in children with pervasive developmental disorder. *Journal of Autism and Developmental Disorders*, *33*(1), 31–45.

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.

Constantino, J. N., Gruber, C. P., Davis, S., Hayes, S., Passanante, N., & Przybeck, T. (2004). The factor structure of autistic traits. *Journal of Child Psychology and Psychiatry*, *45* (4), 719–726.

Critchley, H., Daly, E., Phillips, M., Brammer, M., Bullmore, E., Williams, S., Van Amelsvoort, T., Robertson, D., David, A., & Murphy, D. (2000). Explicit and implicit neural mechanisms for processing of social information from facial expressions: a functional magnetic resonance imaging study. *Human Brain Mapping*, *9*(2), 93–105.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

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

Elliot, C. D. (1990). *Differential abilities scale* (*DAS*). San Antonio, TX: Psychological Corporation.

Gilmour, J., Hill, B., Place, M., & Skuse, D. H. (2004). Social communication deficits in conduct disorder: a clinical and community survey. *Journal of Child Psychology and Psychiatry*, *45*(5), 967–978.

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

Klein-Tasman, B., Risi, S., & Lord, C. (in press). The effect of language and task demands on the diagnostic effectiveness of the Autism Diagnostic Observation Schedule (ADOS): The impact of module choice. *Journal of Autism and Developmental Disorders*.

Klin, A., Jones, W., Schultz, R., Volkmar, F., & Cohen, D. (2002). Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism *Archives of General Psychiatry*, *59*(9), 809–816.

Lord, C., Risi, S., DiLavore, P., Shulman, C., Thurm, A., & Pickles, A. (2006). Autism from two to nine. *Archives of General Psychiatry, 63*(6), 694–701.

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., Pickles, A., & Rutter, M. (2000). The Autism diagnostic observation schedule-generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders*, *30*(3), 205–223.

Lord, C., Rutter, M., DiLavore, P., & Risi, S. (1999). *Autism diagnostic observation schedule: Manual*. Los Angeles: Western Psychological Services.

Mullen, E. (1995). *Mullen scales of early learning* (AGS ed.). Circle Pines, MN: American Guidance Service.

Muthen, L. K., Muthen, B. O. (1998). *M-plus user's guide*. Los Angeles, CA: Muthen and Muthen.

Noterdaeme, M., Mildenberger, K., Sitter, S., & Amorosa, H. (2002). Parent information and direct observation in the diagnosis of pervasive and specific developmental disorders. *Autism*, *6*(2), 159–168.

Risi, S., Lord, C., Gotham, K., Corsello, C., Chrysler, C., Szatmari, P., Cook, E., Leventhal, B., & Pickles, A. (2006). Combining information from multiple sources in the diagnosis of autism spectrum disorders. *Journal of the American Academy of Child and Adolescent Psychiatry, 45*, 1094–1103.

Robertson, J. M., Tanguay, P. E., L'Ecuyer, S., Sims, A., & Waltrip, C. (1999). Domains of social communication handicap in autism spectrum disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, *38*(6), 738–745.

Rogers, S. J., Hepburn, S. L., & Wehner, E. A. (2004). Parent reports of sensory functioning in toddlers with autism and those with other developmental disorders. *Journal of Autism and Developmental Disorders*, *33*(6), 631–642.

Rutter, M., Le Couteur, A., & Lord, C. (2003). *Autism diagnostic interview-revised—WPS* (WPS ed.). Los Angeles: Western Psychological Services.

Schultz, R. T., Gauthier, I., Klin, A., Fulbright, R. K., Anderson, A. W., Volkmar, F., Skudlarski, P., Lacadie, C., Cohen, D. J., & Gore, J. C. (2000). Abnormal ventral temporal cortical activity during face discrimination among individuals with autism and Asperger syndrome. *Archives of General Psychiatry*, *57*, 331–340.

Siegel, B., Vukicevic, J., Elliott, G., & Kraemer, H. (1989). The use of signal detection theory to assess DSM-III-R criteria for autistic disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, *28*, 542–548.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling*. Boca Raton, FL: Chapman & Hall/CRC Press.

Sparrow, S. S., Balla, D. A., & Cicchetti, D. V. (1984). *Vineland adaptive behavior scales*. Circle Pines, MN: American Guidance Service.