Published in final edited form as:
J Autism Dev Disord. 2009 September ; 39(9): 1305–1320. doi:10.1007/s10803-009-0746-z.

The Autism Diagnostic Observation Schedule – Toddler Module: A new module of a standardized diagnostic measure for autism spectrum disorders

Rhiannon Luyster,
University of Michigan Autism and Communication Disorders Center

Katherine Gotham,
University of Michigan Autism and Communication Disorders Center

Whitney Guthrie,
University of Michigan Autism and Communication Disorders Center

Mia Coffing,
University of Michigan Autism and Communication Disorders Center

Rachel Petrak,
University of Michigan Autism and Communication Disorders Center

Karen Pierce,
University of California – San Diego

Somer Bishop,
University of Michigan Autism and Communication Disorders Center

Amy Esler,
University of Michigan Autism and Communication Disorders Center

Vanessa Hus,
University of Michigan Autism and Communication Disorders Center

Rosalind Oti,
University of Michigan Autism and Communication Disorders Center

Jennifer Richler,
University of Michigan Autism and Communication Disorders Center

Susan Risi, and
University of Michigan Autism and Communication Disorders Center

Correspondence concerning this article should be addressed to Rhiannon Luyster, Ph.D., 185 Cambridge St., 6th Floor, Boston MA, 02114. Phone: 617-643-3627. luyster@chgr.mgh.harvard.edu.
Rhiannon Luyster is now at Harvard Medical School, Boston, Massachusetts. Rachel Petrak is now at the University of Michigan School of Public Health, Ann Arbor, Michigan. Somer Bishop is now at the University of Wisconsin, Madison, Wisconsin. Amy Esler and Jennifer Richler are now at the University of Minnesota, Minneapolis, Minnesota. Vanessa Hus is now at the University of Washington, Seattle, Washington.

Some of the data from this paper were previously presented at the 2006 International Meeting for Autism Research (IMFAR) in Montreal, the 2nd World Autism Congress & Exhibition in Cape Town, South Africa, the 2007 Society for Research in Child Development conference in Boston, Massachusetts and at the 2008 IMFAR in London, England.

The Toddler Module (Lord, Luyster, Gotham & Guthrie) is currently in press at Western Psychological Services. The authors of this paper received no royalties from the Toddler Module while it was under development, nor did Drs. Lord or Risi receive royalties for use of any other ADOS modules, due to an agreement with the University of Michigan such that all profits from the authors' use of the measure are donated to charity. The authors of the Toddler Module will receive royalties upon its publication.

Catherine Lord
University of Michigan Autism and Communication Disorders Center

Abstract

The Autism Diagnostic Observation Schedule (ADOS; Lord et al., 2000) is widely accepted as a "gold standard" diagnostic instrument, but it is of restricted utility with very young children. The purpose of the current project was to modify the ADOS for use in children under 30 months of age. A modified ADOS, the ADOS Toddler Module (or Module T), was used in 360 evaluations. Participants included 182 children with best estimate diagnoses of ASD, non-spectrum developmental delay or typical development. A final set of protocol and algorithm items was selected based on their ability to discriminate the diagnostic groups. The traditional algorithm "cutoffs" approach yielded high sensitivity and specificity, and a new range of concern approach was proposed.

Keywords

autism spectrum disorders; diagnosis; ADOS; infants; toddlers

Almost ten years ago, the standardization of a revised Autism Diagnostic Observation Schedule (ADOS), a semi-structured assessment for the diagnosis of autism spectrum disorders (ASD) (Lord, Rutter, DiLavore & Risi, 1999) was described. The ADOS has gradually become an integral part of many research and clinical protocols of children suspected of having an autism spectrum disorder (ASD). Due to the growing understanding of symptoms in the first two years of life and the desire of researchers and clinicians to have standardized instruments for use with infants and young toddlers, there is a need for diagnostic tools that are appropriate for very young children.

This paper presents a new Toddler Module of the ADOS. The Toddler Module retains the original spirit and many of the original tasks of the ADOS, but is intended for use in children under 30 months of age who have nonverbal mental ages of at least 12 months. The scope of this report is to provide a summary of the new measure, the procedures used to develop it, a description of the standardization sample and relevant psychometrics.

In introducing this new module, it is valuable to review the structure of the previously published ADOS. The ADOS evaluates social interaction, communication and play through a series of planned "presses" (Lord et al., 1989) in the context of a naturalistic social interaction. Some of the presses are intended to offer a high level of structure for the participant, while others are intended to provide less structure. All presses, however, afford contexts for both initiations and responses, which are then coded in a standardized manner. An algorithm, which sums the scores of particular items from the measure, yields a classification indicative of autism, ASD or non-spectrum conditions. This classification can then be used by a clinician or researcher as one part of a comprehensive diagnostic process.

The first ADOS was introduced in the late 1980s and was intended for children who had a spoken language age equivalent of at least 36 months. A revision was published in 2000 that reflected the need for the measure to be applicable to a wider range of chronological and developmental ages. The 2000 version provided four separate (but overlapping) modules for individuals of different ages and language abilities. The updated ADOS did indeed extend the usefulness of the original ADOS below a language age of 3 years, but research has indicated that it remains of limited value for children with nonverbal mental ages below 16 months (Gotham, Risi, Pickles & Lord, 2007). For this young population, the ADOS Module 1 algorithm is over-inclusive, meaning that it classifies about 81 percent (19% specificity) of

children with intellectual disabilities and/or language impairments as having autism or ASD when clinical judgment deems that they do not. Revised Module 1 algorithms (Gotham et al., 2007; Gotham et al., 2008) improve specificity but only to 50%.

In recent years, it is precisely this age range, the first two years of life, that has become one of the central concentrations of autism research efforts. Researchers have used creative methodologies to explore the early differences in children who are later diagnosed with ASD, including retrospective videotape analysis, as well as the identification of infants at high risk for ASD (usually the younger siblings of children diagnosed with ASD). The ADOS has been of limited use in these projects, because many of the children fell chronologically or developmentally below the floor of the measure. However, a number of standardized direct observational measures have been developed for use with young children at risk for ASD, such as the Screening Tool for Autism in Two-Year-Olds (STAT; Stone, Coonrod, Turner & Pozdol, 2004), the Communication and Symbolic Behavior Scales Developmental Profile (CSBS-DP; Wetherby, 2001) and the Autism Observational Scale for Infants (AOSI; Bryson, Zwaigenbaum, McDermott, Rombough & Brian, 2008).

Each of these measures serves a different purpose within ASD research. They vary in terms of their target age range, whether they are intended to be used as a screening or diagnostic measure, and whether they were designed to be ASD-specific. The STAT is one of many measures which are intended to be ASD-specific screeners, used in clinics or other specialty centers to identify children at-risk for ASD. It is not intended to be a diagnostic measure, and it is designed for use with children 24 to 35 months of age. The CSBS is intended to be a screening and evaluation measure of communication, social and symbolic abilities in a broad population of children, including children with ASD as well as those with other, non-spectrum conditions. It is designed to identify children (between 12 and 24 months of age) at risk for general developmental delay rather than ASD in particular. However, the use of a specific scoring system (the Systematic Observation of Red Flags) with the CSBS allows for the identification of children at risk for ASD (Wetherby et al., 2004). The AOSI is intended to be an ASD-specific measure, used to detect symptoms of ASD in children between 6 and 18 months of age. Although it may eventually be established as a diagnostic measure, it is not yet proposed to be used as such (Bryson, Zwaigenbaum, McDermott, Rombough, & Brian, 2008). Therefore, although these (and other) important measures for young children have been established, none of them offer a standardized way to reach a diagnostic classification for very young children suspected of having ASD.

A standardized diagnostic measure applicable for infants and young toddlers is also needed for early identification efforts. As public awareness of ASD heightens, parents have been more likely to seek out an evaluation for their very young children. The average age of parental concern is between 15 and 18 months (Chawarska, Paul et al., 2007; DeGiacomo & Fombonne, 1998), and some parents (particularly those who already have one child on the spectrum) have concerns about their child from the earliest months of life. Early identification has been strongly promoted by federal and advocacy organizations with the idea that earlier provision of services will be associated with better outcomes. These findings all point to the need for professionals to be equipped to handle diagnostic assessments for very young children. The Toddler Module should be a useful component of such assessments. One caveat, however, is that diagnostic decisions made very early in life are less stable than those made, for instance, at ages closer to 3 years (Charman et al., 2005; Turner & Stone, 2007). This has been taken into consideration in recommendations for interpretation of the Toddler Module scores (discussed below).

The Toddler Module offers new and modified ADOS activities and scores appropriate for children under 30 months of age who have minimal speech (ranging from no spoken words to simple two-word phrases), have a nonverbal age equivalent of at least 12 months and are

walking independently. Communication, reciprocal social interaction, and emerging object use and/or play skills are all targeted by the module. The ADOS, particularly Modules 1 and 2 – which are intended for developmentally younger children – is designed around the general model that the examiner presents loosely structured and highly motivating materials and activities (e.g., bubbles, snack, remote activated toys) in order to see how the child responds, and whether he/she then makes initiations in order to maintain the interaction.

As in the previously published ADOS modules, each activity of the Toddler Module provides a hierarchy of presses for the examiner. Items that were judged to be appropriate for infants and toddlers were selected from Module 1 of the ADOS and from the PL-ADOS, an early version of the ADOS intended for pre-verbal children (DiLavore, Lord & Rutter, 1995). Additional activities and codes were written based on a review of empirical studies on early development (Behne, Carpenter, Call & Tomasello, 2005; Phillips, Baron-Cohen & Rutter, 1992). Some of the items from previous ADOS versions were re-written to be more appropriate for younger children, and all codes were written on a 4-point scale, ranging from '0' (*no evidence of abnormality related to autism*) to '3' (*definite evidence, such that behavior interferes with interaction*). Eleven activities are included in the Toddler Module (see Table 1), and there are 41 accompanying ratings.

The Toddler Module follows the same basic structure as the Module 1. It should be conducted in a small child-friendly room, and a familiar caregiver should always be present. Simpler cause-and-effect materials are included as well as toys that require the development of more representational and/or imaginative play. Because some of the Module 1 activities – such as a pretend birthday party – may be unfamiliar to younger children, everyday contexts (i.e., a bath-time routine) have been substituted.[1]

Another substantial design change was made because younger children may make fewer explicit and directed initiations towards an unfamiliar adult than older children (Sroufe, 1977). Consequently, in the Toddler Module, we have added instances of the examiner structuring an interaction and waiting for a minimal change in the child's behavior, such as a shift in gaze, facial expression or vocalization. These new activities require less complex motor responses than the Module 1 tasks.

As with other ADOS modules, detailed notes should be taken by the examiner during administration, and coding should be done immediately after the module is complete. Perhaps even more so than other modules, the success and validity of the Toddler Module is dependent on the skill of the examiner. Infants and toddlers, whether typically developing or not, are particularly sensitive to the introduction of new situations and new people (Bohlin & Hagekull, 1993). Indeed, this age range is associated with the development of important components of social and environmental awareness, such as stranger anxiety. As such, the validity of the Toddler Module assumes the clinical skills required to navigate the needs of very young children and carry out the administration and scoring in a reliable fashion.

Design Decisions

Pilot analyses indicated that children chronologically and developmentally younger than 12 months of age consistently obtained elevated scores on early versions of the Toddler Module items, regardless of best estimate diagnostic group. We therefore set a lower cutoff of 12 months nonverbal mental age. In order to allow for the possibility of including children who were performing at age-level, we similarly set the lower cutoff of 12 months chronological age. However, it was anticipated that the final sample would include very few children in the ASD

[1]Inquires about Toddler Module protocols, kits and training should be directed to Western Psychological Services.

group who approached this lower chronological age cutoff, due to the commonality of developmental delays in children with ASD.

It was also necessary to determine at what developmental point children should receive the Module 1, rather than the Toddler Module. Preliminary analyses indicated that Module 1 ADOS sensitivity (percent of children with ASD exceeding the cutoff) and specificity (percent of children without ASD falling below the cutoff) for children over the age of 30 months was superior to the Toddler Module. For this reason, children over 30 months of age were not included in any further analyses, and the methods and results described below exclude these older children. Once a child is over the age of 30 months, he/she should receive the Module 1 of the ADOS (assuming that the child does not yet have sufficient language for a Module 2). Upon mastering spontaneous, non-echoed phrases made up of three independent units, regardless of age, a child should receive the Module 2 of the ADOS.

Diagnostic Algorithm

A subset of items comprise the diagnostic algorithms (see Table 2), following the format of the other modules. Algorithm items are structured according to the domains used in the revised ADOS algorithms (Gotham et al., 2007): Social Affect and Restricted, Repetitive Behaviors. These two domains have been shown to better capture the factor structure of ADOS data than the original three-factor structure (Gotham et al., 2008;Gotham, Risi, Pickles, & Lord, 2007). All items contribute to *one* overall score with a single diagnostic cutoff.

Recent research has indicated that early diagnostic classification within the autism spectrum (making a distinction between the specific diagnoses of autism and pervasive developmental disorder – not otherwise specified, or PDD-NOS) is relatively unstable in young children, even though diagnoses of ASD more broadly versus other, non-spectrum disorders are consistent over time. Lord et al. (2006) reported that 14 percent of children diagnosed with autism at age 2 shifted their diagnosis to PDD-NOS by age 9. Moreover, in children with an age 2 diagnosis of PDD-NOS, 60 percent shifted into an autism classification by age 9. Turner and colleagues (2006), using another sample, reported similar levels of diagnostic uncertainty within the autism spectrum but in the opposite direction, as have other more recent investigations (Kleinman et al., 2008).

Consequently, the Toddler Module includes only two classifications intended for research use: ASD or non-spectrum. Because of the newness of these methods, the relatively small sample sizes, and the care required in interpreting these results, the emphasis for clinical interpretation is on ranges of scores associated with each algorithm. These ranges are associated with the need for clinical monitoring and follow-up (rather than a focus on a cutoff for ASD) and can reflect *little-or-no, mild-to-moderate, or moderate-to-severe* concern.

The purpose of the ADOS algorithm is to provide a classification for the child's current ASD diagnostic status. In the long run, the predictive validity of these scores is extremely important but beyond the scope of this paper and will need additional follow-up data from this and other projects. As with the rest of the ADOS, the algorithm score should never be used as the only source of information in generating a diagnosis. Details about a child's developmental history, parent descriptions and current cognitive, social, language and adaptive functioning across a variety of contexts, as well as the judgment of a skilled clinician, are all necessary for appropriate diagnosis and recommendations (National Research Council, 2001).

Method

Participants

The sample included all children between the ages of 12 and 30 months from three sources: (1) consecutive referrals of children from 12 to 30 months of age from the clinic at the University of Michigan Autism and Communication Disorders Center, (2) children from University of Michigan projects studying early development of children with communication delays and/or at risk for ASD (predominantly younger siblings of children on the autism spectrum), as well as comparison groups of children recruited for these projects and (3) children participating in research at the University of California – San Diego Autism Center of Excellence. "Best estimate" clinical or research diagnoses were assigned based on clinical impressions of a clinical psychologist or an advanced graduate student in psychology (who had received at least two years of supervised ASD-specific assessment and diagnostic experience). Information from a research version of the Autism Diagnostic Interview-Revised (ADI-R, a parent interview; Rutter, Le Couteur & Lord, 2003), modified to be appropriate for toddlers (see Lord, Shulman & DiLavore, 2004) and direct observation (which included the Toddler Module and standardized language and cognitive testing) was available. Thus, clinical diagnosis was not independent of the ADOS. However, algorithms were not derived until after the samples were collected.

The final sample included data from 162 participants at the University of Michigan Autism and Communication Disorders Center; and data from an additional 20 participants from the University of California, San Diego. Preliminary analyses indicated no site differences in age, developmental level or algorithm scores, both within and across diagnostic groups. The project included children with typical development (TD), non-spectrum disorders (NS) and ASD. All individuals with NS and TD did not meet standard ADI-R criteria for ASD (Risi et al., 2006) and received best estimate diagnoses outside the autism spectrum. Non-spectrum participants had a range of diagnoses, including 14 children with expressive language disorders, 5 children with mixed receptive-expressive language disorders, 9 children with non-specific intellectual disability, 4 children with Down syndrome, and 1 child with Fetal Alcohol Syndrome. In addition, one child had been diagnosed with chromosomal abnormalities, one with anxiety disorder not otherwise specified, one with communication disorder – not otherwise specified, and one with phonological disorder. These children were included to demonstrate that the Toddler Module does not consistently identify ASD in children with similar developmental levels as the ASD sample but who did not have ASD. Included in the sample were thirty-five younger siblings of children with ASD, 19 of whom had themselves been given a diagnosis of ASD, 11 of whom were identified as typically developing and 5 of whom had been diagnosed as non-spectrum.

As part of ongoing longitudinal studies, many participants from each site were seen more than once. These children were seen by a familiar clinician for most of their monthly visits but were evaluated by new clinician every six months, who was blind to their previous performance and tentative diagnosis. Altogether, data were used for 182 individuals, who were seen 360 times in total. There was an average of 2.01 (SD = 2.48, range = 1 to 14) assessments per participant. Children in the ASD group were seen between 1 and 14 times (M = 3.24, SD = 3.48), children in the NS group were seen between 1 and 12 times (M = 2.43, SD = 2.86), and typically developing children were seen between 1 and 12 times (M = 1.29, SD = 1.26). For the majority of the validity and reliability analyses reported below, data were analyzed separately for two groups defined by verbal status during the assessment ("verbal" included children whose scores on the item "Overall Level of Language" were '0'=*Regular use utterances of two or more words*; '1'=*Occasional phrases, mostly single words*; or '2'=*At least five single words or word approximations*; "nonverbal" included children whose scores on this item were either '3'=

Less than five but at least one word or word approximation or '8'=*No spontaneous words or word approximations*).

Score distributions differed according to verbal/nonverbal status in children between 21 and 30 months of age. However, distributions of scores for participants younger than 21 months did not systematically vary by verbal/nonverbal status and generally resembled those of nonverbal participants aged 21–30 months. Therefore, the developmental groups were assigned as follows: (1) all children between 12 and 20 months of age as well as nonverbal children between 21 and 30 months of age (hereafter referred to as "12–20/NV21–30"); and (2) verbal children between 21 and 30 months of age ("V21–30"). Data were only used for one time point for each child (the assessment which included cognitive evaluations was selected for inclusion), so that participants were only represented once in each developmental group. However, the same participant could be included once in both groups (i.e., 12–20/NV21–30 and V21–30). There were 136 participants in the 12–20/NV21–30 group (113 children between 12 and 20 months and 23 nonverbal children between 21 and 30 months) and 71 participants in the V21–30 group. This set of data in which each participant was represented only once per group was termed "Unique Participants." In the "Unique Participants" groups, the average chronological age and/or nonverbal mental age were approximately equivalent across the three diagnostic groups. As anticipated, however, there were fewer very young (i.e., under 15 months of age) children in the ASD (n=1) and NS (n=9) groups than in the TD (n=26) group. See Table 3 for sample characteristics.

Analyses were also run for data from all assessments for all participants in order to take advantage of the larger sample size afforded by including repeated measurements. For these analyses, there were 240 visits in 12–20/NV21–30 (194 visits from children between 12 and 20 months and 46 visits from nonverbal children between 21 and 30 months), and 122 visits in V21–30(see Table 4). This set of groups was termed "All Visits." For these analyses, groups were generally not equivalent on measures of mental age and may have been affected by recruitment biases (e.g., non-spectrum children with more ASD-like symptoms were seen more frequently than children with non-spectrum diagnoses and fewer ASD-related behaviors).

For children who had more than one assessment in the last six months of the project, all available data, including research diagnosis history over the most recent months and chart notes, were used by two examiners to generate consensus best estimate "working diagnoses." More weight was given to most recent diagnosis and "blind diagnoses" made by an examiner not familiar with the child. The average age of diagnosis in the 12–20/NV21–30 sample was 16.27 months (SD=2.71) in the typically developing group, 20.26 months (SD=6.12) in the non-spectrum group and 24.68 (SD=5.44) in the ASD group. In the V21–30 sample, the average age of diagnosis was 22.33 months (SD=2.67) for the typically developing group, 25.00 months (SD=5.35) in the non-spectrum group and 25.33 months (SD=2.90) in the ASD group. Each participant received a minimum of one psychometric evaluation using the Mullen Scales of Early Learning (Mullen, 1995), which yielded verbal and nonverbal language age equivalents. For children with repeated assessments, the Mullen was re-administered every six months. All participants were ambulatory (preliminary results indicated that children who were not yet walking had inflated scores on ADOS items), and none had sensory (visual or hearing) impairments or severe motor impairments.

Procedures

The Toddler Module was administered as part of an assessment by clinical research staff and was scored immediately after administration was complete. Over the course of 43 months, 18 different examiners participated in this study. These examiners all had worked with young children on the autism spectrum intensively in either research or clinical settings for at least two years. Included in this group were advanced graduate students who had both extensively

observed and been directly supervised in ASD assessment and diagnosis. All examiners observed and coded numerous Toddler Modules and had attained three consecutive scorings of at least 80% exact agreement with other reliable coders on item-level scores (at least two of which had to be their own administrations) prior to becoming an independent examiner.

Testing was generally administered in a research room, with tables and chairs appropriate for young children. A familiar caregiver was always present in the room. Coding of the Toddler Module was based solely on the behaviors that occurred during the administration of the measure. This included observations of whether a child was "verbal" (i.e., used phrases or at least 5 single words or word approximations). Behaviors that occurred outside the assessment or during administration of another measure were not considered. Consent, which was approved by the University of Michigan Medical School Institutional Review Board for Human Subject Research or the University of California – San Diego Human Subjects Research Protection Program, was given by parents. Families in longitudinal projects received oral feedback and a brief report; participants in other studies received a gift card to a local store.

Inter-rater reliability for the final version of the Toddler Module was formally assessed using 14 administrations from 13 children (one child contributed two administrations). The administrations were independently coded from videotape by each of seven independent, "blind" raters from the original group of 18 examiners. The videos were selected on the basis of the quality of the recording and because the children were not known to the reliability coders. Eight of these participants had best estimate diagnoses of ASD, 3 participants were typically developing, 1 had a diagnosis of mental retardation, and 1 had a diagnosis of Down syndrome.

Results

Test construction and pilot testing

Numerous drafts of the Toddler Module were generated and evaluated, yielding preliminary results and allowing structural decisions about the measure. Proposed items (some of which are included in the final versions and some of which have been eliminated) were used during child assessments and were reviewed and revised during weekly meetings of clinical and research staff. As the project progressed, new codes were added in order to capture additional aspects of child behavior. New examiners and examiners who previously established reliability on ADOS Modules 1 and 2 then established 80 percent agreement in pairs of raters on each item in order to ensure that inter-rater reliability could be obtained by new administrators.

Distributions of scores on each item were generated within cells of children grouped by chronological age, verbal level and diagnosis. Items which appeared to be "too hard" or "too easy" – that is, where typically developing children were often scoring in the '2' to '3' range or where children with ASD were frequently scoring in the '0' to '1' range – were re-written. Additionally, items where the scores fell only between '0' and '2' (that is, few children were scoring in the '3' category) were revised to expand the distribution. Items were eliminated if their distributions, even after revision, did not successfully distinguish among the diagnostic groups (ASD versus typically developing and ASD versus non-spectrum) using one-way ANOVAs. The few exceptions to this criterion were items which were low-incidence but deemed to be clinically significant (e.g., self injurious behavior). When all item revisions were complete, two researchers (blind to child diagnosis) reviewed all relevant the videotaped administrations and/or notes to re-score the revised items according to the final item structure.

In order to determine if there were clinician-related effects on diagnostic decisions, a binary logistic regression was conducted predicting ASD versus non-spectrum best estimate diagnosis. Covariates included the child's age at the time of administration, IQ and Toddler Module algorithm score. Number of years experience working with children on the autism

spectrum was included as a continuous clinician-related predictor (and ranged from 2 years to over 20 years). Results were significant for IQ (β=.05, e^{β}=1.05, p<.05)and algorithm score (β=.54, e^{β}=1.72, p<.01), but not forage (β=.11, e^{β}=1.11, p=ns)or the clinician-related variable (β=.12, e^{β}=1.14, p=ns). Similar results were obtained when using a categorical clinician-related predictor (level of professional education), again indicating that there were not clinician-related effects on best-estimate diagnosis.

Validity Study

The goal of the validity study was to create a modified set of codes and algorithm items that could be used with children between 12 and 30 months of age.

Validity of individual items—Following the item revisions and recoding described above, validity was assessed on a final set of 41 items which either showed markedly different distributions across diagnostic groups or which had high clinical or theoretical importance but rare endorsements. Correlation matrices were generated according to diagnostic group using data from unique participants; these included the complete item set as well as verbal and nonverbal mental age, verbal and nonverbal IQ, and chronological age variables. Items which were highly correlated with each other were identified, and some items were eliminated from consideration for the toddler algorithm in order to reduce collinearity (Note: detailed item data will be available in the Toddler Module manual). The strongest association noted between scores and participant characteristics was between “Overall Level of Non-echoed Language” and verbal IQ (r=−.71 across diagnostic groups, n=113), so no items were excluded on this basis.

Exploratory factor analyses were then conducted in Mplus (Muthen & Muthen, 1998) with a focus on ASD participants only. Due to the small sample size, these analyses were not intended to identify a latent class structure for the item data, but rather to provide an assessment of the potential influence of cognitive level and chronological age on these data. Chronological age ceased to load onto any factor when the sample was divided into the two developmental groups (12–20/NV21–30 and V21–30). Verbal mental age did not load onto any factor for either developmental group.

Validity of algorithm—In order to select items for the algorithm, item means and standard deviations were generated across diagnostic groups. The items that best differentiated between diagnoses for the “Unique Participants” and “All Visits” subsets within narrow age/language groups (which were eventually collapsed into the 12–20/NV21–30 and V21–30 groups)were identified. Similarities in diagnostically differential items across the younger (under 21 months) and nonverbal groups, as well as a distinct “best” item set for older verbal toddlers, confirmed the validity of the two developmental groupings used for these analyses. A pool of 17 items was identified as strong candidates for a new Toddler Module algorithm based on their differential distributions across diagnostic group and their relatively low correlations with each other and with chronological age and IQ. Some of these items were new items in the Toddler Module and others had been included in previous Module 1 ADOS algorithms.

Next, best items for each developmental group were summed to generate trial algorithms specifically for the 12–20/NV21–30 and V21–30 groups. Visits missing data from more than 2 algorithm items were excluded from these analyses. Scores of ‘2’ and ‘3’ were collapsed in candidate items following the ADOS convention intended to prevent any one item from exerting undue influence on the total score, and conversely, a score of ‘1’ on the Unusual Eye Contact item was converted to ‘2’ on the algorithms in order to reflect the importance of even subtle differences in eye contact. Receiver Operating Characteristic (ROC) curves (Siegel, Vukicevic, Elliott & Kraemer, 1989) allow sensitivity and specificity percentages to be

generated for each total score in a scale. For 12–20/NV21–30 visits, sensitivity and specificity was generated for both trial toddler algorithms as well as the ADOS Module 1, No Words algorithm for "Unique Participants" and "All Visits" subsets of data. For V21–30 visits, ROC curve analyses were run for both trial toddler algorithms and the ADOS Module 1, Some Words algorithm for both "Unique Participants" and "All Visits" subsets. Specificity was evaluated in comparisons of ASD versus non-spectrum participants, and again for ASD versus non-spectrum and typical cases combined, for all possible cutoffs in each of the three possible algorithms. These algorithms were then re-tested by systematically omitting items to ensure that each item contributed to the final differentiations. Within each developmental group, the strongest algorithm out of the three tested was selected by identifying the cutoff score that maximized both sensitivity and specificity across "Unique Participants" and "All Visits" subsets, and that maintained specificity in ASD versus non-spectrum distinctions as well as ASD versus non-spectrum and typical combined. The results are shown in Table 5.

For children under 21 months and nonverbal toddlers, the same set of items that comprise the ADOS Module 1, No Words algorithm also maximized predictive validity of this measure, though it is important to note that codes and scores associated with items of the same name in the Toddler Module and Module 1 are not identical. A cutoff of 12 on this 12–20/NV21–30 algorithm yielded 91% sensitivity and 91% specificity for ASD versus non-spectrum comparisons of unique participants. This cutoff also maintained sensitivity values at 87% or greater and specificity at 86% or greater when applied to "All Visits" and comparisons of typically developing children (see Table 5 for details). Moreover, the cutoff performed similarly when applied to the 12–20 and NV21–30 groups separately, using both Unique Participants and All Visits samples. All sensitivity and specificity values exceeded 85%, with one exception (75% specificity for ASD versus non-spectrum in the NV21–30 group, based on a cell size of 8).

For verbal toddlers between 21 and 30 months of age, a new algorithm was superior to the Module 1, Some Words algorithm. As shown in Table 5, a cutoff of 10 on this V21–30 algorithm yielded sensitivity of 88% and specificity of 91% in the ASD versus non-spectrum unique participants. Sensitivity was maintained at 81% or greater and specificity at 83% or greater for all other comparisons, with the lowest in these ranges pertaining to "All Visits" repeated comparisons of ASD and Non-spectrum participants. The V21–30 algorithm is comparable in structure to the ADOS revised algorithms, with 14 items organized into Social Affect (SA) and Restricted, Repetitive Behaviors (RRB) domains (see Table 2 for a list of items by domain). In the new V21–30 algorithm, however, only three of these items describe RRBs versus four RRB items in the 12–20/NV21–30 and other revised algorithms across ADOS modules. This difference in maximum RRB total score between the 12–20/NV21–30 and V21–30 algorithms was not theoretically motivated but rather reflects the selection of items that maximized predictive value of the new algorithms in these developmental groups.

To improve clinical utility of this measure, ranges of concern were identified for the new V21–30 algorithm and the 12–20/NV21–30 algorithm used with young or nonverbal toddlers. Using the "Unique Participants" data, three ranges of concern were set for each algorithm, such that at least 95% of children with ASD and no more than about 10% of typically developing children would fall in the two groups suggesting clinical concern (mild-to-moderate and moderate-to-severe). See Table 6 for results.

For both developmental groups, 82% of children with non-ASD developmental delays were accurately assigned to the little-or-no concern range.

Internal consistency of algorithm—In the new V21–30 algorithm, item-total correlations for "All Visits" ranged from .49 ("Response to Name") to .82 ("Quality of Social Overtures")

for the Social Affect domain, and from .18 ("Hand and Finger Mannerisms") to .42 ("Unusual Sensory Interest in Play Material/Person") for the three items comprising the RRB domain (the third being "Unusually Repetitive Interests or Stereotyped Behaviors," r=.37). Lower correlations within the RRB domain were expected given the heterogeneous nature of these items. Cronbach's alpha was .90 for the SA domain and .50 for the RRB domain, indicating strong and acceptable internal consistency respectively. Correlations between domain totals and participant characteristics (e.g., chronological age, gender, mental age, and IQ) were evaluated within the "Unique Participants" subset only, because of the known effects of recruitment on the composition of the "All Visits" sample. In the older group of verbal toddlers, domains were correlated at .64 with each other. Across all domain total correlations, noneexceeded −.55 with participant characteristics (between verbal IQ and SA total). Correlations with chronological age did not exceed .48 (with SA total), those with mental age did not exceed −.42 (verbal mental age with SA total), and those with nonverbal IQ did not exceed −.51 (with RRB total).

For the younger or nonverbal children receiving 12–20/NV21–30 algorithm, item-total correlations for "All Visits" ranged from .35 ("Gestures") to .81 ("Quality of Social Overtures") in the SA domain, and from .14 ("Hand and Finger Mannerisms") to .44 ("Unusually Repetitive Interests or Stereotyped Behaviors") for the four-item RRB domain. Internal consistency was similar to the older, verbal group findings, with a Cronbach's alpha of .88 for the SA domain and .50 for the RRB domain. For "Unique Participants" in this developmental group, the domains were correlated at .57 with each other. Across all domain total correlations with participant characteristics, noneexceeded-.58(SA total with verbal mental age). Correlations between domain totals and chronological age did not exceed .34 (with SA total). Correlations with nonverbal mental age did not exceed −.17 (with SA total), those with verbal IQ did not exceed −.38 (with SA total), and those with nonverbal IQ did not exceed −.49 (with SA total).

For both algorithms, SA and RRB domain total scores for "Unique Participants" were significantly higher for the ASD sample than the non-spectrum or typically developing groups (see Table 7). Domain totals for the two non-ASD diagnostic groups did not differ significantly, with the exception of SA scores (non-spectrum mean exceeded typically developing) in the 12–20/NV21–30 group. One-way ANOVA and Tukey test statistics are available from the authors.

Reliability Study

Inter-rater reliability of individual items—For reliability analyses, scores indicating that the item was not applicable (generally these were language-related items) were converted to zeros, as is done for algorithm use in the other ADOS modules. Three items (out of a total of 41 items) were either rare or considered particularly valuable in interpreting child behavior ("Stereotyped/Idiosyncratic Use of Words or Phrases," "Self-Injurious Behavior," and "Overactivity")had percent agreements exceeding 90 percent but received such a limited range of scores that they were not included in further reliability analyses.

STATA software (StataCorp, 2007) was used to generate weighted kappas for non-unique pairs of raters (i.e., 28 pairs). Kappas between .4 and .74 were considered good, and kappas at or above .75 were considered excellent (Fleiss, 1986). Out of 38 items, 30 weighted kappas were equal to or exceeded .60 (Mk_w = .67). The remainder exceeded .45.

Inter-rater item reliability for all items in the protocol was assessed by domain by exploring the percent of exact agreement. Because having reliable '3' scores allows better documentation of variation (which is important in treatment studies), the initial set of analyses retained all scores of '0' to'3'. Percent agreement between 70% and 79% was considered fair, 80% to 89% was considered good and above 90% was considered excellent (Cicchetti, Volkmar, Klin, &

Showalter, 1995). For items on the Toddler Module, even using the extended range of '0' to '3' (which reduces agreement), mean exact (percent) agreement was 84%across all items and rater pairs. Thirty of 41 items had exact agreement at or above 80%, and every item received at least 71%agreement across raters. When considered by domain, agreement for codes related to language and communication was generally good: only three items had reliability that was fair(71%, 74% and 75%). Codes related to reciprocal social interaction were mostly good-to-excellent, with only six items falling in the fair range (75% to 78%). Play and restricted, repetitive behaviors had only one item each in the fair range (78% and 75%, respectively), with all others above 80%. All items in the nonspecific behaviors domain had good or excellent inter-rater reliability.

Because the diagnostic algorithm collapses codes of 2s and 3s (to avoid overly weighting any single item in the overall diagnosis), a second set of exact agreement analyses were conducted, collapsing codes of 2 and 3. Mean exact agreement was 87%. Thirty-five of 41 items had exact agreement above 80%, and no item agreement fell below 71%.

Inter-rater reliability of domain scores and algorithm classifications—Intraclass correlations (ICCs) were computed for protocol total scores, as well as algorithm domain and total scores. Calculations were made using both the 12–20/NV21–30 and V21–30 algorithms. ICCs were as follows: protocol total scores = .96; 12–20/NV21–30 algorithm total = .90; V21–30 algorithm total = .99; 12–20/NV21–30 algorithm SA total = .84;V21–30 algorithm SA total = .99; 12–20/NV21–30 algorithm RRB total = .93; V21–30 algorithm RRB total = .74.

Inter-rater agreement in diagnostic classification using a single cutoff of 12 (i.e., ASD or non-spectrum) was 97% on the 12–20/NV21–30 algorithm. Using the V21–30 algorithm with a single cutoff of 10, inter-rater agreement across diagnostic classifications (i.e., ASD or non-spectrum) was 87%. Inter-rater agreement using the three ranges on the 12–20/NV21–30 algorithm (little-or-no concern: scores less than 10, mild-to-moderate concern: scores of 10 to 13, moderate-to-severe concern: scores of 14 and above) was 70%. On the V21–30 algorithm (little-or-no concern: scores less than 8, mild-to-moderate concern: scores of 8 to 11, moderate-to-severe concern: scores of 12 and above), inter-rater agreement for ranges of concern was 87%.

Test-retest reliability—Test-retest reliability was analyzed using data from all children (n=39)who had two Toddler Module administrations within 2 months. Reliability was evaluated using algorithm subtotal scores across the SA and RRB domains, as well as algorithm total scores. Analyses addressing the 12–20/NV21–30 algorithm, which included 31 participants, yielded high test-retest ICCs for the SA total (.83), the RRB total (.75), and the algorithm total score (.86). The mean *absolute* difference across the two evaluations was 0.90 points (*SD* = 3.14) for SA, 0.39 points (*SD* = 1.54) for RRBs and 1.29 points (*SD* = 3.55) for the algorithm total score. Out of the 31 children, 24(77%)were classified consistently across the two evaluations (using the single cutoff of 10 on the algorithm). Out of the 7 participants who shifted between non-spectrum and ASD classification, 3 initially missed the cutoff and then met the cutoff on the second evaluation, while 4 moved from meeting the cutoff to failing to meet. Using the three ranges of concern, 23 (74%) children were classified within the same range across evaluations. Of the 8 participants who shifted between ranges of concern, 1 shifted from the greater level of concern to the lesser one. Seven shifted from little-or-no concern to a concern range or vice-versa (2 from little-or-no concern to mild-to-moderate concern, 4 from mild-to-moderate concern to no concern, and 1 from moderate-to-severe concern to little-or-no concern).

Data for 8 participants who received the V21–30 algorithm twice within two months indicated similarly high ICCs for the SA total (.94), the RRB total (.60), and the algorithm total score (.

95). There was a mean absolute difference across the two evaluations of 0.63 points(*SD* = 2.13) for algorithm total scores, 0.38 points (*SD* = 2.77)for the SA total, and 0.25 points (*SD* = 1.04) for the RRB total. Using the single cutoff of 10, 2 children shifted classifications across evaluations (1 shifting from meeting cutoffs to failing to meet, the other vice-versa) and 6 retained the same classification. Similarly, 5 out of the 8 children remained in the same range of concern across both administrations. Of the remaining 3 children, 1 increased from mild-to-moderate to moderate-to-severe concern, 1 moved from mild-to-moderate to little-or-no concern and the other shifted from little-or-noconcern to mild-to-moderate concern.

Discussion

The Toddler Module contributes a new module to the existing ADOS and permits the use of this standardized instrument with children under 30 months of age. It includes three core areas of observation, namely, language and communication, reciprocal social interaction, play and stereotyped/restricted behaviors or interests. Algorithm scores have acceptable internal consistency and excellent inter-rater and test-retest reliability. The algorithm, using both the formal cutoff and the ranges of concern, has excellent diagnostic validity for ASD versus non-spectrum conditions. Children who receive the Toddler Module should have a nonverbal age equivalent of at least 12 months and be walking independently. If a child has not yet attained all of these milestones, Toddler Module results may be elevated due to developmental factors and must be interpreted with care.

The lower chronological age limit for the Toddler Module is proposed to be 12 months. This is estimated based on the nonverbal mental age requirement of 12 months and the increased observation of more children on the spectrum performing at age expectations (Chakrabarti & Fombonne, 2001). However, the current sample included only one child under the age of 15 months who met this nonverbal mental age criteria. Therefore, the present investigation validated the proposed algorithms only down to 15 months of chronological age. It is clear that the Toddler Module tasks and items are appropriate for children in the age range. It is also apparent that in chronological ages under 15 months, the algorithm had good specificity in this sample (due to the higher numbers of 12 to 15 month olds in the non-spectrum and typically developing groups). However, the sensitivity of the proposed algorithm has not yet been established for children with ASD who (a) have nonverbal mental ages of at least 12 months and (b) are between 12 and 15 months of chronological age. This will need to be addressed in future investigations in order for the lower chronological age cutoff to be confirmed.

As with other modules of the ADOS, the Toddler Module algorithm should be interpreted cautiously and in conjunction with other sources of information. Use of the algorithm ranges should be one element of a comprehensive diagnostic assessment, in which the final diagnostic decision must be made using the best judgment of the clinician. This is particularly important when evaluating very young children, for whom the lines of typical and atypical development can be very unclear and for whom behavior can change over a few months. Moreover, differential diagnosis can be especially challenging in toddlers because symptoms may emerge gradually. An attempt has been made to structure the Toddler Module algorithm in a manner which – as much as is possible – accommodates these observations by generating ranges of concern rather than strict classifications. In addition, because research has indicated that early specific ASD diagnoses (autism and PDD-NOS versus ASD) have questionable stability in younger populations, the algorithms provide only one research cutoff for all ASD.

The single cutoffs proposed for the new algorithms should be interpreted in a fashion consistent with the ADOS: “an individual who meets or exceeds the cutoffs … has scored within the range of a high proportion of participants with [ASD] who have similar levels of expressive language and deficits in social behavior and in the use of speech and gesture as part of social

interaction" (Lord et al., 2000, p. 220). However, in order to warrant an ASD diagnosis, the individual must otherwise exhibit behaviors consistent with the criteria as outlined in formal diagnostic criteria (American Psychiatric Association, 1994). That is, it is possible for a child to meet a cutoff and not receive a formal diagnosis of ASD according to clinical judgment. Conversely, it is also possible for a child to score below the cutoff and for a clinician to judge that the child does meet formal criteria for an ASD diagnosis. Some aspects of the algorithm scores (i.e., negative association with early verbal scores) highlight the importance of thoughtful clinical interpretation of algorithm results, because certain features of the child which are non-specific to ASD (like early language delay) may elevate scores. Because verbal ability in this study was defined by MSEL (Mullen, 1995) scores, and – as with other measures – the early MSEL scores are heavily biased to social communication (e.g., "recognizes own name" and "plays gesture/language game"), the correlations between Toddler Module scores and early verbal ability scores seemed inevitable, though a clearer separation between ADOS scores and eventual language ability would be ideal.

The ranges of concern which are incorporated into the algorithm are intended to reflect the diagnostic uncertainty that is often faced when evaluating very young children, whether because of developmental variability or confounding conditions (such as global developmental delay or early language impairment). Nevertheless, by expanding the number of categories from two diagnostic groupings (ASD and non-spectrum) to three ranges of concern (little-or-no, mild-to-moderate, moderate-to-severe), more variation would be expected. Thus, the ranges are intended primarily as "sign-posts" along a continuous range of scores that show excellent stability in intra-class correlations, across raters and re-assessments several months later. Scores falling into the little-or-no concern range suggest that the child demonstrates no more behaviors associated with ASD than children in this age range who do not have ASD. Generally, scores which fall into the mild-to-moderate range should be considered an indicator of behaviors likely to be consistent with an ASD. Children whose scores fall into this range should receive further ASD-specific evaluation and follow-up in the next several months, including ongoing monitoring of cognitive and language development, as well as ASD symptoms. Note that a minority of children with non-spectrum conditions and typical development also scored in this range, so there is considerable heterogeneity within it. In contrast, algorithm scores falling into the moderate-to-severe range of concern were strongly consistent with an eventual diagnosis of ASD (with only 3–6% false positives). Regardless, whether using the research-oriented cutoff or the clinically-oriented ranges of concern, the onus is on the examiner to interpret behaviors and scores within the broader developmental and assessment context. In cases of diagnostic uncertainty, it is important to be clear with parents (particularly of very young children) about the importance of ongoing monitoring of child development and thorough follow-up.

The importance of the algorithm and its items may lead ADOS administrators to ask why additional codes are necessary. There are two primary purposes for including codes in the ADOS which are not algorithm items. First, the present investigation is an initial attempt to generate a research and clinical tool. New information from larger, independent investigations may result in improved algorithms using a different set of items, as has been the case for the ADOS (Gotham et al., 2007; Gotham et al., 2008). Second, the non-algorithm items describe important aspects of ASD and may characterize the strengths and weaknesses of individual children. Changes in non-algorithm items may provide valuable information concerning response to treatment and, more speculatively, different etiological subtypes or patterns of behavior.

The young age of the children receiving the Toddler Module means that the examiner may face some additional issues in interpreting ADOS results. Specifically, some infants and toddlers may be very uncomfortable in the evaluation context, where they are faced with an unknown

adult, unfamiliar toys, and a novel clinic or laboratory setting. The examiner must, therefore, gauge whether behavior observed in the ADOS context is representative of behavior in other settings. This is especially important if something about the ADOS assessment – an unskilled examiner, the absence of a familiar caregiver, cultural differences in expected child behavior – might suggest that the child's behavior is "off". Fortunately, because the Toddler Module requires that (barring unique circumstances, such as children recently placed in foster care) a familiar caregiver is always present in the room, the examiner should get feedback from the caregiver about whether the child's behavior during the ADOS was representative of day-to-day interactions. If something about the ADOS administration indicates that the observation did not capture the child's every-day behavior, the scores should be interpreted accordingly and more information should be sought through a home observation or a repeated assessment.

In addition to the above child-related factors, there are important examiner-related factors which must be considered when using the Todder Module. All examiners in the present investigation had at least two years of intensive experience working with young children at risk for and identified with ASD. Furthermore, all examiners had participated extensively – either through consensus discussions or supervision – in generating early differential diagnoses. This high level of experience in working with the relevant population is extremely important, in terms of both clinical skill and the validity of clinical judgment. Although the current study did not find an association between degree of clinical experience and final diagnostic judgments, previous projects have reported that limited experience is associated with lower clinician agreement for specific spectrum diagnoses (Stone, Lee, Ashford, & Brissie, 1999). As previously stated, information obtained from the Toddler Module should be only one component of a diagnostic decision. Nevertheless, it is extremely important that the measure be used by individuals who have sufficient clinical experience to appropriately interpret the observations and algorithm results.

Results and observations from the Toddler Module may be useful beyond the diagnostic context. Parents, intervention providers and teachers often report that the strengths and difficulties noted during the administration can help in understanding an individual child and developing programming goals. Therefore, clinicians should make a concerted effort to thoroughly explain the key observations in behavioral terms (rather than simply in terms of scores and cutoffs), describing which behaviors were noted and which were less consistent or absent. When appropriate, examiners should generate suitable recommendations based on the ADOS observations which can be applied to educational and treatment plans at home and at school.

The predictive validity of very early diagnosis (under 30 months) is a question currently being addressed by many investigators (Chawarska, Klin, Paul & Volkmar, 2007; Landa & Garrett-Mayer, 2006; Wetherby et al., 2004; Zwaigenbaum et al., 2005). The focus of the Toddler Module development is to provide a standardized method of quantifying descriptions of behaviors that correspond to experienced clinicians' best estimate clinical diagnosis of ASD at a given point in time. The Toddler Module provides information with good to excellent internal consistency and inter-rater reliability for items, domains and research diagnostic categories. Stability across raters within clinical ranges was good for older, verbal children but less good for the nonverbal and younger children. Across time, about three-quarters of children remained in the same clinical range of concern for both algorithms, and slightly fewer remained in the same diagnostic category. Thus, variations both in rater and in time do make a difference in a child's outcome on the Toddler Module. Follow-up studies of the long-term predictive value of these scores will be critical in determining the extent to which they, and other early measures of diagnostic risk, predict outcome and response to treatment. In the meantime, consideration of scores as continuous dimensions and as one marker (along with other measures) of relative risk of ASD and need for follow-up seems most appropriate. In research,

the diagnostic categories may help in standardizing assessments across studies and establishing replicable criteria for study inclusion. Again, however, algorithm classification should be considered in the context of other information.

There are some limitations to the present investigation. The sample size is small and did not permit very fine-grained age groupings. Of particular importance is the limited number of NS children in the 21–30 V group and the limited number of very young children with ASD. Furthermore, the ASD sample was considerably larger than the comparison samples, which may have affected the sensitivity and specificity of the cut-offs. It was also noted that many of the children in the ASD sample had age-level nonverbal abilities. Although a higher-functioning sample (versus a more impaired one) may better approximate the cohort of children currently receiving diagnoses (Chakrabarti & Fombonne, 2001), it provides less information about symptom overlap between ASD and other non-spectrum conditions in children with marked intellectual disabilities. All of these factors may have affected the observed results and need to be addressed in additional samples to confirm the validity of the currently proposed measure guidelines (e.g., use for children under 15 months of age) and algorithm construction.

Test-retest reliability was evaluated over the course of up to 2 months (rather than over the course of several days) and may be confounded by developmental changes. In addition, it is important to acknowledge that evaluation and diagnosis were not completely independent processes because the administration of the ADOS was part of standard practice, although diagnosis was independent of algorithm results. Finally, a cross-validation sample is required to test the algorithm cut-offs (and their associated sensitivity and specificity). It will be important to address these concerns, as well as broader questions such as calibration (using algorithm scores as continuous measurements of severity) through replication in future independent studies.

In sum, the Toddler Module is a new, standardized module intended to extend the application of the ADOS to children as young as 12 months of age who have nonverbal mental ages of at least 12 months. It is appropriate for use with children up to the age of 30 months or until children acquire phrase speech. Replication of the psychometric results reported here with larger, more diverse samples of children with early-appearing, non-spectrum conditions as well as with ASD is crucial, as are follow-up studies that provide information about predictive validity. We hope that researchers and clinicians alike find it a useful tool in supporting families and children with autism spectrum disorders and advancing our understanding of these conditions.

Acknowledgments

This work was supported by NRSA F31MH73210-02 from the National Institute of Mental Health to Rhiannon Luyster, as well as grants MH57167 and MH066469 from the National Institute of Mental Health and HD 35482-01 from the National Institute of Child Health and Human Development, and funding from the Simons Foundation to Catherine Lord. Support was also provided by a grant from the Department of Education to Amy Wetherby. We thank Andrea Cohan, Christina Corsello, Pamela Dixon Thomas, Lee Anne Green Snyder, Alexandra Hessenius, Marisela Huerta, Lindsay Jackson, Jennifer Kleinke, Fiona Miller, Rebecca Niehus and Dorothy Ramos for their assistance in data collection. We would also like to express our gratitude to the families and children in the Toddlers study, the Word Learning project, and the First Words project.

References

American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4. Washington, DC: Author; 1994.

Behne T, Carpenter M, Call J, Tomasello M. Unwilling versus unable: Infants' understanding of intentional action. Developmental Psychology 2005;41(2):328–337. [PubMed: 15769189]

Bohlin G, Hagekull B. Stranger wariness and sociability in the early years. Infant Behavior and Development 1993;16(1):53–67.

Bryson SE, Zwaigenbaum L, McDermott C, Rombough V, Brian J. The Autism Observational Scale for Infants: Scale development and reliability data. Journal of Autism and Developmental Disorders 2008;38(4):731–738. [PubMed: 17874180]

Chakrabarti S, Fombonne E. Pervasive developmental disorders in preschool children. Journal of the American Medical Association 2001;285(24):3093–3099. [PubMed: 11427137]

Charman T, Taylor E, Drew A, Cockerill H, Brown J, Baird G. Outcome at 7 years of children diagnosed with autism at age 2: Predictive validity of assessments conducted at 2 and 3 years of age and pattern of symptom change over time. Journal of Child Psychology and Psychiatry 2005;46(5):500–513. [PubMed: 15845130]

Chawarska K, Klin A, Paul R, Volkmar F. Autism spectrum disorder in the second year: Stability and change in syndrome expression. Journal of Child Psychology and Psychiatry 2007;48(2):128–138. [PubMed: 17300551]

Chawarska K, Paul R, Klin A, Hannigen S, Dichtel LE, Volkmar F. Parental recognition of developmental problems in toddlers with autism spectrum disorders. Journal of Autism and Developmental Disorders 2007;37(1):62–72. [PubMed: 17195921]

Cicchetti D, Volkmar F, Klin A, Showalter D. Diagnosing autism using ICD-10 criteria: A comparison of neural networks and standard multivariate procedures. Child Neuropsychology 1995;1:26–37.

DeGiacomo A, Fombonne E. Parental recognition of developmental abnormalities in autism. European Journal of Child and Adolescent Psychiatry 1998;7:131–136.

DiLavore P, Lord C, Rutter M. Pre-linguistic Autism Diagnostic Observation Schedule (PLADOS). Journal of Autism and Developmental Disorders 1995;25(4):355–379. [PubMed: 7592249]

Fleiss, J. Reliability of measurements. In: Fleiss, J., editor. The design and analysis of clinical experiments. New York: John Wiley & Sons; 1986. p. 2-31.

Gotham K, Risi S, Pickles A, Lord C. The Autism Diagnostic Observation Schedule (ADOS): Revised algorithms for improved diagnostic validity. Journal of Autism and Developmental Disorders 2007;37(4):613–627. [PubMed: 17180459]

Gotham K, Risi S, Dawson G, Tager-Flusberg H, Joseph R, Carter A, Hepburn S, McMahon W, Rodier P, Hyman SL, Sigman M, Rogers S, Landa R, Spence MA, Osann K, Flodman P, Volkmar F, Hollander E, Buxbaum J, Pickles A, Lord C. A replication of the Autism Diagnostic Observation Schedule (ADOS) revised algorithms. Journal of the American Academy of Child and Adolescent Psychiatry 2008;47(6):642–651. [PubMed: 18434924]

Kleinman JM, Ventola PE, Pandey J, Verbalis AD, Barton M, Hodgson S, Green J, Dumont-Mathieu T, Robins DL, Fein D. Diagnostic stability in very young children with autism spectrum disorders. Journal of Autism and Developmental Disorders 2008;38(4):606–615. [PubMed: 17924183]

Landa R, Garrett-Mayer E. Development in infants with autism spectrum disorders: A prospective study. Journal of Child Psychology and Psychiatry 2006;47(6):629–638. [PubMed: 16712640]

Lord C, Risi S, DiLavore P, Shulman C, Thurm A, Pickles A. Autism from two to nine. Archives of General Psychiatry 2006;63(6):694–701. [PubMed: 16754843]

Lord C, Risi S, Lambrecht L, Cook EH, Leventhal BL, DiLavore P, Pickles A, Rutter M. 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 2000;30(3):205–223. [PubMed: 11055457]

Lord, C.; Rutter, M.; DiLavore, P.; Risi, S. Autism Diagnostic Observation Schedule (ADOS). Los Angeles: Western Psychological Services; 1999.

Lord C, Rutter M, Goode S, Heemsbergen J, Jordan H, Mawhood L, Schopler E. Autism Diagnostic Observation Schedule: A standardized observation of communicative and social behavior. Journal of Autism and Developmental Disorders 1989;19:185–212. [PubMed: 2745388]

Lord C, Shulman C, DiLavore P. Regression and word loss in autism spectrum disorder. Journal of Child Psychology and Psychiatry 2004;45(5):936–955. [PubMed: 15225337]

Mullen, E. Mullen Scales of Early Learning. Circle Pines, MN: American Guidance Service, Inc; 1995.

Muthen, LK.; Muthen, BO. M-plus user's guide. Los Angeles: Muthen and Muthen; 1998.

National Research Council. Educating children with autism. Washington, DC: National Academy Press; 2001.

Phillips W, Baron-Cohen S, Rutter M. The role of eye contact in goal detection: Evidence from normal infants and children with autism or mental handicap. Development and Psychopathology 1992;4:375–383.

Risi S, Lord C, Gotham K, Corsello C, Chrysler C, Szatmari P, Cook EH Jr, Leventhal BL, Pickles A. Combining information from multiple sources in the diagnosis of autism spectrum disorders. Journal of the American Academy of Child and Adolescent Psychiatry 2006;45(9):1094. [PubMed: 16926617]

Rutter, M.; Le Couteur, A.; Lord, C. Autism Diagnosic Interview -Revised (ADI-R). Los Angeles: Western Psychological Services; 2003.

Siegel B, Vukicevic J, Elliott G, Kraemer H. 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 1989;28:542–548. [PubMed: 2768150]

Sroufe LA. Wariness of strangers and the study of infant development. Child Development 1977;48:1184–1199.

StataCorp. Stata Statistical Software: Release 10. College Station, TX: StataCorp LP; 2007.

Stone W, Coonrod EE, Turner LM, Pozdol SL. Psychometric properties of the STAT for early autism screening. Journal of Autism and Developmental Disorders 2004;34(6):691–701. [PubMed: 15679188]

Stone W, Lee E, Ashford L, Brissie J. Can Autism be Diagnosed Accurately in Children Under Three Years? Journal of Child Psychology & Psychiatry 1999;20(2):219–226. [PubMed: 10188704]

Turner L, Stone W. Variability in outcome for children with an ASD diagnosis at age 2. Journal of Child Psychology and Psychiatry 2007;48(8):793–802. [PubMed: 17683451]

Turner L, Stone WL, Pozdol S, Coonrod EE. Follow-up of children with autism spectrum disorders from age 2 to age 9. Autism 2006;10(3):243–265. [PubMed: 16682397]

Wetherby, A. Communication and Symbolic Behavior Scales Developmental Profile, Preliminary Normed Edition. Baltimore, MD: Paul H. Brookes Publishing; 2001.

Wetherby A, Woods J, Allen L, Cleary J, Dickinson H, Lord C. Early indicators of autism spectrum disorders in the second year of life. Journal of Autism and Developmental Disorders 2004;34(5): 473–493. [PubMed: 15628603]

Zwaigenbaum L, Bryson S, Rogers T, Roberts W, Brian J, Szatmari P. Behavioral manifestations of autism in the first year of life. International Journal of Developmental Neuroscience 2005;23:143–152. [PubMed: 15749241]

Table 1

Toddler Module Activities

1a. Free Play
1b. Free Play -- Ball
2. Blocking Toy Play
3. Response to Name
4a. Bubble Play
4b. Bubble Play -- Teasing Toy Play
5a. Anticipation of a Routine with Objects
5b. Anticipation of a Routine with Objects -- Unable Toy Play
6. Anticipation of a Social Routine
7. Response to Joint Attention
8. Responsive Social Smile
9a. Bathtime
9b. Bathtime -- Ignore
10. Functional & Symbolic Imitation
11. Snack

Table 2

Algorithm Items

12–20/NV21–30 (12–20 months & Non-verbal 21–30 months)	V21–30 (Verbal 21–30 months)
Social Affect	
Frequency of Spontaneous Vocalization Directed to Others	Response to Name
Gestures	Ignore
Shared Enjoyment in Interaction	Requesting
Showing	Pointing
Unusual Eye Contact	Unusual Eye Contact
Facial Expressions Directed to Others	Facial Expressions Directed to Others
Integration of Gaze and Other Behaviors During Social Overtures	Integration of Gaze and Other Behaviors During Social Overtures
Spontaneous Initiation of Joint Attention	Spontaneous Initiation of Joint Attention
Response to Joint Attention	Amount of Social Overtures/Maintenance of Attention: CAREGIVER
Quality of Social Overtures	Quality of Social Overtures
	Overall Quality of Rapport
Restricted, Repetitive Behaviors	
Unusual Sensory Interest in Play Material/Person	Unusual Sensory Interest in Play Material/Person
Hand and Finger Movements/Posturing	Hand and Finger Movements/Posturing
Unusually Repetitive Interests or Stereotyped Behaviors	Unusually Repetitive Interests or Stereotyped Behaviors
Intonation of Vocalizations or Verbalizations	

Note. 12–20/NV21–30 algorithm is a modified version of the Revised Module 1, No Words algorithm from 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(4), 613–627.

Table 3

Description of "Unique Participants" sample in validity analyses

	12–20 All			Nonverbal 21–30		Verbal 21–30		
	ASD	**Non-spectrum**	**Typical**	**ASD**	**Non-spectrum**	**ASD**	**Non-spectrum**	**Typical**
***N* (male, female)**	20 (19, 1)	24 (22, 2)	68 (47, 21)	15 (12, 3)	8 (3, 5)	24 (21, 3)	11 (10, 1)	36 (24, 12)
Mean Chronological age (SD)	17.70^{a} (1.90)	15.33^{b} (2.48)	15.82^{b} (2.69)	22.20^{a} (1.86)	25.38^{b} (2.26)	26.00^{a} (2.62)	23.45^{b} (2.66)	22.28^{b} (1.11)
Chronological age minimim-maximum	13–20	12–19	12–20	21–28	22–28	22–30	21–29	21–25
Mean Nonverbal mental age (SD)	18.05 (2.98)	16.94 (3.10)	18.62 (3.18)	19.87 (3.61)	19.62 (7.11)	25.04 (3.56)	24.82 (5.78)	25.40 (3.32)
Nonverbal mental age minimim-maximum	13–23	12–24	12–26	12–26	12–35	19–33	17–37	20–35
Mean Verbal mental age (SD)	13.05^{a} (4.15)	13.35^{a} (2.96)	16.93^{b} (4.03)	11.03 (5.03)	13.13 (6.58)	21.79 (5.12)	22.32 (7.06)	25.39 (3.94)
Verbal mental age minimim-maximum	6–21	9–19	10–26	1–19	4–26	12–31	12–38	19–36

Note. ASD=autism spectrum disorder; SD=standard deviation. All ages are in months. There were no nonverbal typically developing children between 21 and 30 months.

$^{a, b, c}$Superscript letters refer to post hoc Scheffe test: When two groups in the same row and column are marked with different letters, they are significantly different ($p < .01$) from each other; groups marked with the same letter/without superscript letters are not significantly different.

Table 4

Description of "All Cases" sample in validity analyses

	12–20 All			Nonverbal 21–30		Verbal 21–30		
	ASD	**Non-spectrum**	**Typical**	**ASD**	**Non-spectrum**	**ASD**	**Non-spectrum**	**Typical**
***N* (male, female)**	55 (53, 2)	47 (45, 2)	90 (61, 29)	32 (27, 5)	14 (9, 5)	59 (56, 3)	24 (23, 1)	39 (26.13)
Mean Chronological age (SD)	17.20 (2.34)	15.87 (2.48)	15.93 (2.69)	23.31 (2.39)	24.71 (2.81)	25.54^{a} (2.71)	24.79^{a} (2.78)	22.46^{b} (1.37)
Chronological age minimim-maximum	12–20	12–20	12–20	21–29	21–30	21–30	21–29	21–27
Mean Nonverbal mental age (SD)	17.21$^{a, b}$ (2.92)	16.65^{a} (4.33)	18.63^{b} (3.34)	20.59 (3.28)	19.43 (5.26)	24.49 (3.53)	23.75 (4.46)	25.46 (3.84)
Nonverbal mental age minimim-maximum	12–23	12–37	12–28	12–26	12–35	18–33	17–37	17–35
Mean Verbal mental age (SD)	11.87^{a} (3.79)	13.39^{a} (5.37)	17.04^{b} (4.18)	12.61 (4.99)	13.00 (5.17)	21.41^{a} (5.93)	21.35$^{a, b}$ (5.83)	25.18^{b} (4.61)
Verbal mental age minimim-maximum	6–21	6–38	10–26	1–25	4–26	11–31	12–38	10–36

Note. ASD=autism spectrum disorder; SD=standard deviation. All ages are in months. There were no nonverbal typically developing children between 21 and 30 months.

$^{a, b, c}$Superscript letters refer to post hoc Scheffe test: When two groups in the same row and column are marked with different letters, they are significantly different ($p < .01$) from each other; groups marked with the same letter/without superscript letters are not significantly different.

Table 5

Sensitivity and specificity of the algorithm cutoffs used with the ADOS-Toddler Module

"Unique Participants" sample							
12–20/NV21–30 algorithm[a] (12–20 months & Non-verbal 21–30 months)				**V21–30 algorithm[b] (Verbal 21–30 months)**			
ASD vs NS		**ASD vs NS, TD**		**ASD vs NS**		**ASD vs NS, TD**	
(n=35)	**(n=34)**	**(n=35)**	**(n=101)**	**(n=24)**	**(n=11)**	**(n=24)**	**(n=47)**
Sens	**Spec**	**Sens**	**Spec**	**Sens**	**Spec**	**Sens**	**Spec**
91%	91%	91%	94%	88%	91%	88%	94%
"All Cases" sample							
12–20/NV21–30 algorithm[a] (12–20 months & Non-verbal 21–30 months)				V21–30 algorithm[b] (Verbal 21–30 months)			
ASD vs NS		ASD vs NS, TD		ASD vs NS		ASD vs NS, TD	
(n=87)	(n=64)	(n=87)	(n=153)	(n=59)	(n=24)	(n=59)	(n=63)
Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec
87%	86%	87%	91%	81%	83%	81%	90%

Note. ASD=autism spectrum disorder; NS=non-spectrum; TD=typically developing; Sens=Sensitivity; Spec=Specificity.

[a] Using a cutoff of 12.

[b] Using a cutoff of 10.

Table 6

Percent of participants falling into ranges of concern by diagnostic group

12–20/NV21–30 (12–20 months & Non-verbal 21–30 months)				V21–30 (Verbal 21–30 months)			
Range	**ASD (n=35)**	**NS (n=34)**	**TD (n=67)**	**Range**	**ASD (n=24)**	**NS (n=11)**	**TD (n=36)**
Little-or-no concern Scores: 0–9	3	82	89	Little-or-no concern Scores: 0–7	4	82	92
Mild-to-moderate concern Scores: 10–13	20	12	8	Mild-to-moderate concern Scores: 8–11	12	18	8
Moderate-to-severe concern Scores: 14+	77	6	3	Moderate-to-severe concern Scores: 12+	84	--	--

Note. ASD=autism spectrum disorder; NS=non-spectrum; TD=typically developing.

Table 7

Mean algorithm domain scores by diagnostic group

(12–20 months & Non-verbal 21–30 months)12–20/NV21–30				V21–30 (Verbal 21–30 months)			
Domain	**ASD (n=35)**	**NS (n=34)**	**TD (n=67)**	**Domain**	**ASD (n=24)**	**NS (n=11)**	**TD (n=36)**
Social Affect	13.8 (4.1)	6.2 (3.9)	3.7 (2.4)	Social Affect	12.1 (4.6)	3.1 (2.7)	2.3 (2.1)
Restricted, Repetitive Behaviors	4.3 (1.4)	1.5 (1.4)	1.6 (1.3)	Restricted, Repetitive Behaviors	3.6 (2.0)	0.6 (1.0)	0.9 (1.0)

Note. ASD=autism spectrum disorder; NS=non-spectrum; TD=typically developing. Standard deviations in parentheses.