Journal of Autism and Developmental Disorders, Vol. 30, No. 3, 2000

The Autism Diagnostic Observation Schedule–Generic: A Standard Measure of Social and Communication Deficits Associated with the Spectrum of Autism

Catherine Lord,[1] Susan Risi,[1] Linda Lambrecht,[1] Edwin H. Cook, Jr.,[1] Bennett L. Leventhal,[1] Pamela C. DiLavore,[2] Andrew Pickles,[3] and Michael Rutter[4]

The Autism Diagnostic Observation Schedule–Generic (ADOS-G) is a semistructured, standardized assessment of social interaction, communication, play, and imaginative use of materials for individuals suspected of having autism spectrum disorders. The observational schedule consists of four 30-minute modules, each designed to be administered to different individuals according to their level of expressive language. Psychometric data are presented for 223 children and adults with Autistic Disorder (autism), Pervasive Developmental Disorder Not Otherwise Specified (PDDNOS) or nonspectrum diagnoses. Within each module, diagnostic groups were equivalent on expressive language level. Results indicate substantial interrater and test–retest reliability for individual items, excellent interrater reliability within domains and excellent internal consistency. Comparisons of means indicated consistent differentiation of autism and PDDNOS from nonspectrum individuals, with some, but less consistent, differentiation of autism from PDDNOS. A priori operationalization of DSM-IV/ICD-10 criteria, factor analyses, and ROC curves were used to generate diagnostic algorithms with thresholds set for autism and broader autism spectrum/PDD. Algorithm sensitivities and specificities for autism and PDDNOS relative to nonspectrum disorders were excellent, with moderate differentiation of autism from PDDNOS.

KEY WORDS: Autism Diagnostic Observation Schedule; PDDNOS; non-autistic-spectrum diagnoses; expressive language skill.

INTRODUCTION

The Autism Diagnostic Observation Schedule-Generic (ADOS-G) is a semistructured assessment of social interaction, communication, play, and imaginative use of materials for individuals who may have autism or other pervasive developmental disorders (PDDs). As part of the schedule, planned social occasions, referred to as "presses" (Lord *et al.*, 1989; Murray, 1938), are created in which a range of social initiations and responses is likely to appear. In the same way, communication opportunities are designed to elicit a range of interchanges. Play situations are included to allow observation of a range of imaginative activities and social role-play. The goal of the ADOS-G is to provide presses that elicit spontaneous behaviors in standardized contexts. Structured activities and materials, and less structured interactions, provide standard contexts within the ADOS-G in which social, communicative, and other behaviors relevant to the understanding of PDDs are observed.

The ADOS-G is the direct outgrowth of two similar diagnostic instruments: the Autism Diagnostic Observation Schedule (ADOS; Lord *et al.*, 1989) and the Pre-Linguistic Autism Diagnostic Observation Scale (PL-ADOS; DiLavore, Lord, & Rutter, 1995). The

[1] University of Chicago, Chicago, Illinois.
[2] University of North Carolina, Chapel Hill, North Carolina.
[3] University of Manchester, Manchester, England.
[4] Social Genetic and Developmental Psychiatry Research Centre. Institute of Psychiatry, London, England.

0162-3257/00/0600-0205$18.00/0

ADOS was first introduced in the 1980s as a method of standardizing direct observations of social behavior, communication, and play in children suspected of having autism. It used immediate coding; videotapes of the schedule offered the potential for more detailed analyses later. Ideas for activities and for the behaviors to be coded during the schedule were adapted from empirical research in autism and child development. The ADOS was intended to be administered to children between the ages of 5 and 12, who had expressive language skills at least at the 3-year-old level. It was proposed as a complementary instrument to the Autism Diagnostic Interview (ADI; Le Couteur *et al.*, 1989), an investigator-based parent or caregiver interview that yielded a description of history, as well as current functioning, in areas of development related to autism. The instruments were developed primarily for research diagnosis of autism over a range of cognitive levels from moderate mental retardation to normal intelligence, with training required on each.

Two factors led to modifications in the original ADOS and ADI, which resulted in the creation of the PL-ADOS (DiLavore *et al.*, 1995) and the Autism Diagnostic Interview-Revised (ADI-R; Lord, Rutter, & Le Couteur, 1994). One factor was the growing interest in using the instruments in clinical settings. Because children under age 5 now constitute the bulk of referrals for a first diagnosis of autism, there was a need to extend the age and verbal limits of the ADOS and the ADI to be appropriate for younger and nonverbal children. The second factor was the authors' participation in a longitudinal study of children referred for possible autism at the age of 2. These studies served as an impetus to modify the ADOS and ADI in such a way that the instruments addressed the concerns of parents and fit the abilities of children functioning at infant and toddler levels (DiLavore *et al.*, 1995; Lord *et al.*, 1994).

It became apparent that the ADOS conversational style and the context of sitting at a table for 30 minutes were not effective for eliciting a range of social-communicative behavior or play from very young children. Accordingly, although concepts, principles, and general strategies from the ADOS could be maintained, more flexible, briefer activities and greater use of play materials were necessary. The result was the PL-ADOS, an observational schedule for nonverbal young children that served as a downward extension for the ADOS, rather than a replacement.

The PL-ADOS was effective in discriminating 2- to 5 year-old-children with autism from children with non-autism spectrum developmental delays (DiLavore *et al.*, 1995). However, it tended to be underinclusive for children with autism who had some expressive language. Thus, a tool was required to address the needs of children who fell between the PL-ADOS and ADOS in language skills. Furthermore, the ADOS consisted primarily of activities intended for school-age children. Additional or alternative tasks were needed for adolescents and adults. Experience with the ADOS and PL-ADOS also indicated a number of ways in which both instruments could be more efficient and reliable. The ADOS-G was designed in response to these factors.

The ADOS-G differs from the preceding instruments in several ways. It is aimed at providing standard contexts for the observation of behavior for a broader developmental and age range of individuals suspected of having autism. The schedule now consists of four modules. Each one is appropriate for children and adults at different developmental and language levels, ranging from no expressive or receptive use of words, to fluent, complex language in an adult. Only one module, lasting about 30 minutes, is administered to any individual at a given point in time.

Expressive language level is probably the strongest predictor of outcome in autism spectrum disorders, at least in individuals beyond the preschool level (Kobayashi, Murata, & Yoshinaga, 1992; Venter, Lord, & Schopler, 1992). Because expressive language level affects almost every aspect of social interaction and play, it has been particularly difficult to disentangle the effects of language level from severity of autism in verbal individuals with autism spectrum disorders (Happé, 1995; Mahoney *et al.*, 1998). Research has shown that children with mental retardation, with or without autism, appear more socially competent, less anxious, and more flexible when language demands are low relative to their level of ability (Mesibov, Schopler, & Hearsey, 1994). In previous versions of the ADOS and PL-ADOS, this resulted in overdiagnosis of autism in children with insufficient language ability for the tasks and underdiagnosis of autism in children whose language abilities exceeded those for which the scale was intended (e.g., children with phrase speech who were given the PL-ADOS; DiLavore *et al.*, 1995; Lord *et al.*, 1989). The introduction of the different modules in the ADOS-G is intended to minimize the possible biasing effect of variations in language skill by offering different tasks and codings in the appropriate modules. In the ADOS-G, the examiner uses the module that best matches the expressive language skills of the individual child or adult in order to make judgments about social and

communicative abilities as independent as possible from the effects of absolute level of language delay.

The modules provide social-communicative sequences that combine a series of unstructured and structured situations. Each situation provides a hierarchy of presses for particular social behaviors. Module 1, based on the PL-ADOS, is intended for children who do not use spontaneous phrase speech consistently. As shown in Table I, it consists of 10 activities with 29 accompanying ratings. Module 2 is intended for children with some flexible phrase speech who are not verbally fluent. It consists of 14 activities with 28 accompanying ratings. Module 3 provides 13 activities and 28 ratings. It is based on the ADOS and is intended for verbally fluent children for whom playing with toys is age-appropriate. The operational definition of verbal fluency is the spontaneous, flexible use of sentences with multiple clauses that describe logical connections within a sentence. It requires the ability to talk about objects or events not immediately present. Module 4 contains the socioemotional questions of the ADOS, along with interview items about daily living and additional tasks. It is intended for verbally fluent adults and for adolescents who are not interested in playing with toys such as action figures (usually over 12–16 years). This module consists of 10–15 activities with 31 accompanying ratings. The difference between Modules 3 and 4 lies primarily in whether information about social-communication is acquired during play or through a conversational interview. It is important to note that adolescents or adults may feel uncomfortable when presented with the toys for young children that are available in Modules 1 and 2. An experimental version of the ADOS-G, appropriate for minimally verbal or nonverbal adolescents or adults is under development by the authors.

Modules 1 and 2 are often conducted while moving around a room, reflecting the interests and activity levels of young children or children with very limited language; Modules 3 and 4 generally take place sitting at a table and involve more conversation and language without a physical context. Though the tasks and materials in the different modules vary, the general principles involving the deliberate variation of the examiner's be-

Table I. Modules 1–4: Activities[a]

Module 1 Preverbal/ single words/ simple phrases	Module 2 Flexible phrase speech	Module 3 Fluent speech child/adolescent	Module 4 Fluent speech adolescent/adult
Anticipation of a social routine	Construction task	Construction task	Construction task[a]
	Make-believe play	Make-believe play	Current work/school/daily living[b]
Functional and symbolic imitation	Joint interactive play	Joint interactive play	Socioemotional questions: Plans and dreams
Free play	Free play	Break	Break
Snack	Snack		
Response to name	Response to name	Cartoons	Cartoons[a]
Response to joint attention	Response to joint attention	Socioemotional questions: Emotions	Socioemotional questions: Emotions
Birthday party	Birthday party	Socioemotional questions: Friends/loneliness/marriage	Socioemotional questions: Friends/loneliness/marriage
Bubble play	Bubble play	Socioemotional questions: Social difficulties/annoyance	Socioemotional questions: Social difficulties/annoyance
Anticipation of a routine with objects	Anticipation of a routine with objects	Creating a story	Creating a story
	Demonstration task	Demonstration task	Demonstration task
	Conversation	Conversation/reporting a nonroutine event	Conversation/reporting a nonroutine event
	Description of picture	Description of picture	Description of picture[a]
	Looking at a book	Telling a story from a book	Telling a story from a book

[a] Activities in a line are similar in intent whenever possible, but not always.
[b] Indicates an optional activity.

havior using a hierarchy of structured and unstructured social behaviors are the same. Unlike the ADOS or the PL-ADOS, all codings on the ADOS-G are made after the schedule is administered; notes are taken during specific tasks, but scoring encompasses the entire schedule. Some tasks were eliminated from earlier versions if they did not add unique information. Items are coded for imaginative play and restricted or repetitive behaviors or interests, but these items are not included in the diagnostic algorithms, because the single, relatively brief observation period does not provide an optimal opportunity for their assessment.

The ADOS-G (Lord, Rutter, DiLavore, & Risi, 1999) also differs from the earlier ADOS (Lord *et al.*, 1989) and PL-ADOS (DiLavore *et al.*, 1995) in its attempt to include a broader range of individuals in the population it addresses. Unlike the earlier instruments, which included only participants who met standard criteria for autism or who showed other delays but no evidence of autism spectrum disorder, the ADOS (Lord *et al.*, 1999) standardization includes groups of children and adults for each module with diagnoses of PDDNOS. The long-term goal is to devise a diagnostic tool that measures social and communication deficits in numerous autism spectrum disorders (ASD), including PDDNOS, Asperger syndrome, childhood disintegrative disorder, and atypical autism. PDDNOS was selected as the most easily identifiable group in our clinic population that could potentially be matched on level of expressive language to our clinic sample of individuals with autism. Individuals with other types of PDDs were not included. Throughout the paper, the term PDDNOS is used to refer to the specific diagnoses required for our sample; autism spectrum disorder (ASD) is used to refer to the general conceptualization of a continuum of social and communication deficits associated with autism.

Use of the ADOS-G is clearly related to the skill of the examiner. It requires practice in administering the activities, scoring, and observation. Within a clinic or research group, before they can be regarded as competent to use the instruments with clinical or research populations, examiners are expected to obtain interrater reliability with each other and with consensus ratings on videotapes provided by the authors before using the instruments. For research purposes, examiners are expected to attend standardized training workshops conducted by workshop leaders who have demonstrated their reliability during and following their own training, and to obtain reliability with workshop leaders as well as within their each research site. The ADOS-G is now available in a published version as the ADOS-WPS edition (Lord, *et al.*, 1999), including a manual with detailed statistics and a test kit.

Diagnostic Algorithms and Use of the ADOS-G for Classification

Subsets of items in each module are used to generate separate diagnostic algorithms for each module in the ADOS-G. Items in these algorithms are listed in Table II. Items and the thresholds for classification of autism and of autism spectrum disorder differ for each module in the ADOS-G. However, the general principles and procedures for computation are the same across modules and similar to DSM-IV (American Psychiatric Association, 1994) and ICD-10 (World Health Organization, 1993). Classification is made on the basis of exceeding thresholds on each of two domains: social behavior and communication, and exceeding a threshold for a combined social-communication total. The ADOS-G is intended to be one source of information used in making a diagnosis of autism spectrum disorders, but is not sufficient to do so on its own. Because only a small window of time is considered, the ADOS-G does not offer an adequate opportunity to measure restricted and repetitive behaviors (though such behaviors are coded if they occur). Thus, ADOS-G algorithms include only items coding social behaviors and communication. Because it consists of codings made from a single observation, the ADOS-G does not include information about history or functioning in other contexts. This means that the ADOS-G alone cannot be used to make complete standard diagnoses. For example, to receive a DSM-IV or ICD-10 diagnosis of autism, an individual must show evidence of restricted or repetitive behaviors and evidence of abnormalities manifest before 36 months. If an individual showed restricted, repetitive behaviors on the ADOS-G, this would increase the likelihood of a diagnosis of autism, but other historical information, such as provided by the ADI-R, would be required.

The original intention of the ADOS-G was to provide separate classification, using different algorithms for autism and PDDNOS as a specific disorder, with an emphasis upon qualitative differences between different disorders. Samples were selected such that participants' diagnoses were likely to be clearly differentiated by direct observation during the ADOS-G. Thus, participants were required to have clinical diagnoses of autism or PDDNOS or no evidence of any ASD, based on overall best estimates. However, when verbal level was controlled, after many analyses, there was no suggestion of consistent qualitative differences across

Table II. Modules 1–4: Algorithm and Other Items for Diagnosis of Autism DSM-IV/ICD-10 for Social and Communication Domains[a]

Module 1 Preverbal/ single words/ simple phrases	Module 2 Flexible phrase speech	Module 3 Fluent speech child/adolescent	Module 4 Fluent speech adolescent/adult
		Algorithm items	
Stereotyped/idiosyncratic words or phrases	Stereotyped/idiosyncratic words or phrases	Stereotyped/idiosyncratic use of words or phrases	Stereotyped/idiosyncratic use of words or phrases
Gestures	Descriptive, conventional, instrumental gestures	Descriptive, conventional, instrumental gestures	Descriptive, conventional, instrumental gestures
Unusual eye contact	Unusual eye contact	Unusual eye contact	Unusual eye contact
Facial expressions directed to others	Facial expressions directed to others	Facial expressions directed to others	Facial expressions directed to others
Quality of social overtures	Quality of social overtures	Quality of social overtures	Quality of social overtures
Response to joint attention[2]	Amount of reciprocal social communication	Amount of reciprocal social communication	Amount of reciprocal social communication
Shared enjoyment[2,3,4]	Quality of social response	Quality of social response	Quality of social response
Use of other's body to communicate	Conversation	Conversation	Conversation
Pointing	Pointing to express interest		Emphatic or emotional gestures
Showing[2]	Overall quality of rapport	Overall quality of rapport	
Frequency of vocalization directed to others	Amount of social overtures	Insight[4]	Empathy/comments on others' emotions[3]
Spontaneous initiation of joint attention	Spontaneous initiation of joint attention	Reporting of events[4]	Responsibility
		Other Items	
Immediate echoing	Immediate echoing	Immediate echoing	Immediate echoing
Speech abnormalities	Speech abnormalities	Speech abnormalities	Speech abnormalities
Imagination/functional play	Imagination/functional play	Imagination	Imagination
Mannerisms	Mannerisms	Mannerisms	Mannerisms
Unusual sensory behaviors	Unusual sensory behaviors	Unusual sensory behaviors	Unusual sensory behaviors
Repetitive interests and behaviors	Repetitive interests and behaviors	Excessive, specific interests	Excessive, specific interests
		Rituals and compulsive behaviors	Rituals and compulsive behaviors
Overactivity	Overactivity	Overactivity	Overactivity
Negative behavior	Negative behavior	Negative behavior	Negative behavior
Anxiety	Anxiety	Anxiety	Anxiety

[a] Whenever possible, items on the same horizontal line reflect similar or identical codes but, for the sake of space, this is not always true. Superscripts indicate items that appear in other modules, but are not in other algorithms. Nonalgorithm items that only occur in one or two modules are not included here.

modules in the behavioral patterns of participants with autism and PDDNOS. As shown below in Results, the distributions of item and domain scores from autism to PDDNOS were continuous, with no evidence of clustering, other than minor associations that were unique to specific modules. Because these findings are consistent with other studies, the decision was made to use a single set of algorithm items for each module. Autism is then placed on a continuous dimension with ASD, including disorders within the autism spectrum that do not meet criteria for autism (i.e., PDDNOS), and then "other" disorders, using the algorithm scores. The term "other" was selected to indicate that this is a grouping made only by exclusion. Thus, the ADOS-G algorithms discriminate between the narrower definitions of autism and broader definitions of ASD, including PDDNOS, on the basis of severity. Severity includes the number of symptoms as well as the severity of each symptom. The implications of this method of scoring in the ADOS-G are discussed in more detail later.

An ADOS-G autism classification requires meeting or exceeding each of the three thresholds (social, communication, social-communication total) for autism. If thresholds for autism are not met, an ADOS-G classification of ASD is appropriate when the three ASD thresholds are met or exceeded. The ASD thresholds are, in all cases, lower than those of autism. As discussed above, it is important to distinguish between an ADOS-G classification and an overall diagnosis. An overall autism diagnosis requires abnormalities in restricted, repetitive behaviors and early manifestations of the disorder. Thus, there may be cases in which an individual receives an ADOS-G classification of autism, but a clinical diagnosis of autism, PDDNOS, or Asperger disorder. Conversely, a clinical diagnosis of PDDNOS may be made in the presence of significant social abnormalities and restricted, repetitive behaviors, without communication dysfunction; in this case, the behavior of an individual might meet criteria for only the social domain and so not receive an ADOS-G classification of ASD but still receive an overall diagnosis of PDDNOS. These discrepancies illustrate the importance of combining information from the ADOS-G with history and parent report, such as in the ADI-R, and clinical judgment in integrating the information from different sources.

METHOD

Participants

General Issues in Subject Characteristics and Selection Across Modules

The initial sample for all modules consisted of 381 consecutive referrals to the Developmental Disorders Clinic at The University of Chicago. Consensus clinical diagnoses were assigned based on clinical impressions of a clinical psychologist and a child psychiatrist who each interviewed the parents and observed the child separately and discussed discrepant impressions until they reached a "best estimate" diagnosis. The clinicians had access to history, results of a physical examination, and scores on the Autism Diagnostic Interview-Revised (ADI-R; Lord *et al.,* 1994). Direct observations of the individual participant occurred during the ADOS-G, physical exam, psychological testing, and free time with the parents. All individuals receiving autism diagnoses met ADI-R criteria for autism. Because standards for the use of the ADI-R in diagnosis of nonautism PDDs are not established, descriptive information from the ADI-R, but not algorithms, was used in diagnosis of PDDNOS. Agreement for clinical diagnoses made independently was monitored at least once a week throughout the study and remained consistently over 90% for autism and nonspectrum disorders and over 80% for PDDNOS.

From these samples 20–30 participants were selected to be included in the reliability analyses in each module. About one half of the participants in the reliability analyses in each module had autism, one third had PDDNOS, and one sixth had nonspectrum disorders. Because reliability analyses were carried out during the first part of data collection, selection for the reliability studies was made on the basis of order of recruitment and availability of two independent scorings

Additional participants for the validity study were recruited in order to obtain three samples in each module (autistic, PDDNOS, and nonspectrum) that were of adequate size and equivalent verbal mental age (for Modules 1, 2, and 3) or verbal IQ (Module 4). Participants in the three groups were chosen (within a module) to be as close as possible across diagnostic groups in chronological age, gender, and ethnicity. Research centers in which investigators had completed training on the ADOS-G, including Yale (8 participants), University of California San Diego (5 participants), and Newcastle upon Tyne (3 participants), contributed clinical and psychometric information about children and adults who were used to complete the matched samples. Videotapes of ADOS-Gs from these centers were rescored by local staff. Psychometric data, ADI-R scores, and a clinical diagnosis were provided for each of these participants by the originating center.

All individuals with nonspectrum disorders (NS) who were recruited through the clinic were selected because they failed to meet autism criteria on the ADI-R and received independent diagnoses outside the autism spectrum. However, children and adults were predominantly recruited from nonautism clinics within the Department of Psychiatry at The University of Chicago, as well as from local special education programs and group homes. For these individuals, the Autism Screening Questionnaire (ASQ; Berument *et al.,* 2000) was completed, and only individuals with scores below 15 (the suggested cutoff for autism) were included in the samples (all scores were actually 8 or below). Diagnoses of NS participants included mental retardation (often with behavior disorders), receptive-expressive language disorder (often with behavior disorders), attention-deficit hyperactivity disorder and/or oppositional defiant disorder (often with learning disabilities), anxiety disorder, major depression, obsessive-compulsive disorder (the last three diagnoses in Module 4 only), and typically developing children and adults. For each module,

a few typically developing children or adults were recruited to ensure that the ADOS-G did not misclassify individuals without known pathology and to provide intellectual matches for the highest functioning individuals with autism. The number of each nonspectrum group consisting of typically developing individuals ranged from 3 (Modules 1–3) to 7 (Module 4). The nonspectrum groups were not intended to form homogeneous diagnostic groups or to represent any particular group of nonspectrum disorders. The purpose of this group's inclusion was to show that the items and algorithms of the ADOS-G do not routinely identify autism/ASD in individuals of comparable language skill who do not have clinical diagnoses of autism spectrum disorders. Because of the need to recruit nonspectrum participants as specific language matches, research staff who scored the ADOS-G were blind to diagnosis in most, but not all, cases.

Each individual participating in the study received at least one psychometric test yielding an age equivalent in language skills and one test yielding a nonverbal age equivalent. When they were available, scores on the Peabody Picture Vocabulary Test–Third Edition (Dunn & Dunn, 1997) or Peabody Picture Vocabulary Test–Revised (Dunn & Dunn, 1981) and the Raven's Progressive Matrices (Raven, 1956, 1960) were used. These tests were administered to all participants who were seen solely for the purposes of this study and many of the other participants. When these results were not available, scores were used from the Mullen Scales of Early Learning, averaging the two verbal and two nonverbal subtests separately and excluding the gross motor scale (Mullen, 1995), the Differential Ability Scales (Elliott, 1990), the Wechsler Intelligence Scale-Third Edition (Wechsler, 1991) or the Wechsler Adult Intelligence Scale–Revised (Wechsler, 1984). Within modules, distributions of specific tests were similar across diagnostic groups. However, across modules, tests obviously differed according to the participants' level of functioning and chronological age. Participants in verbally equivalent diagnostic groups were not individually matched but selected from within the same age and ability ranges in order to yield equivalent mean scores, with standard deviations as similar as possible.

Ethnicity was relatively comparable across modules and across groups, with 80% Caucasian, 11% African American, 4% Hispanic, 2% Asian American, and 2% children and adults of other or mixed races, participating in the study. These samples were not intended to be representative of a particular population. All participants were native English speakers; none had hearing or visual impairments other than mild visual difficulties corrected with glasses. All participants were ambulatory. None had motor problems more severe than very mild cerebral palsy, which occurred in one nonspectrum, mildly retarded adolescent in Module 4. No participants had identifiable syndromes, except for one boy with Williams syndrome in the nonspectrum group in Module 3.

Module 1. As shown in Table III, 54 children, constituting three groups equivalent in verbal mental age, were selected to be in the validity sample (MAUT = matched autistic group; PDD = PDDNOS; NS = nonspectrum). Twenty-nine of these children (14 MAUT, 7 PDD, 8 NS) were included in the reliability sample, selected on the basis of characteristics described earlier. Children ranged in chronological age from 15 months to 10 years. All children walked independently; none were yet using spontaneous, meaningful three-word phrases, and many had no spoken language during the ADOS-G administration. Many of the children with autism who were initially recruited as participants for Module 1 could not be included in the samples above because they could not be matched to other diagnostic groups on language level. However, because of the clinical importance of documenting autism spectrum disorders in very young, severely delayed children with autism, it was felt that it was worth including a description of these children, even if the degree to which the ADOS-G can discriminate children with and without autism at young developmental ages is not possible to determine. Thus, an additional group of 20 lower functioning children with autism (LAUT) was included. This group was selected to be as close as possible in chronological age to the children with PDDNOS and nonspectrum disorders and equivalent in nonverbal mental age to the NS group. The LAUT group had significantly lower language skills than any of the other groups, F (3, 70) = 25.59, $p < .001$; $\chi^2 > 11.4$, $p < .001$ for Scheffé tests (Scheffé, 1953). There were no other significant differences among the groups in chronological age, verbal mental age, or nonverbal mental age. In total, there were 57 males and 17 females, with males exceeding females by at least a ratio of 2:1 in all groups.

Module 2. As shown in Table III, 55 children were selected for the validity analyses of Module 2, constituting three groups equivalent in verbal mental age. Twenty-three children (9 AUT, 8 PDD, 6 NS) were included in reliability analyses. All children used at least some spontaneous, meaningful three-word utterances, but did not yet meet the criteria for verbal fluency. They ranged in age from 2 to 7 years. For Modules 2, 3, and 4, the following abbreviations are used to identify groups: AUT = autism; PDD = PDDNOS; and

Table III. Description of Subjects in Validity Analyses: Means and Standard Deviations for Modules 1–4[a]

	Module 1			
	Lower autism	Matched autism	PDDNOS	Nonspectrum
N (male, female)	20(16,4)	20(18,2)	17(11,6)	17(12,5)
Chronological age	4.02(1.2)	4.94(1.58)	4.28(1.95)	3.51(2.14)
Verbal mental age	0.98(0.19)	2.21(0.39)	2.20(0.84)	1.93(0.44)
Nonverbal mental age	2.36(0.57)	3.17(1.03)	3.07(1.37)	2.41(0.95)
	Module 2			
	Autism	PDDNOS	Nonspectrum	
N (male, female)	21(15,6)	18(15,3)	16(9,7)	
Chronological age	4.56(1.26)	4.38(1.23)	3.78(1.04)	
Verbal mental age	2.97(0.51)	2.95(0.57)	2.85(0.85)	
Nonverbal mental age	3.94(1.12)	3.79(0.77)	3.15(1.03)	
	Module 3			
	Autism	PDDNOS	Nonspectrum	
N (male, female)	21(19,2)	20(17,3)	18(11,7)	
Chronological age	9.14(2.36)	7.26(1.23)	10.09(4.94)	
Verbal mental age	6.93(1.80)	6.94(1.86)	6.94(1.89)	
Nonverbal mental age	8.00(2.07)	6.84(1.99)	7.78(2.22)	
	Module 4			
	Autism	PDDNOS	Nonspectrum	
N (male, female)	16(14,2)	14(11,3)	15(12,3)	
Chronological age	18.65(7.79)	21.59(8.56)	19.11(6.27)	
Verbal IQ	99.94(22.29)	105.5(21.46)	99.73(26.69)	
Nonverbal IQ	94.06(28.22)	105.21(21.82)	103.8(27.48)	

[a] All ages are in years. Scores for chronological age, verbal mental age, verbal IQ, and nonverbal IQ are means. Standard deviations, where applicable, are in parentheses.

NS = nonspectrum. There were no significant differences among diagnostic groups in chronological age, verbal mental age, or nonverbal mental age.

Module 3. As shown in Table III, 59 children and adolescents, constituting three groups, were selected to be equivalent in verbal mental age for Module 3. Twenty-six children and adolescents were included in the reliability analyses (12 AUT, 6 PDD, 8 NS). All participants in this sample met the criteria for verbal fluency specified earlier. They ranged in age from 3 to 20 years. The PDDNOS group was significantly younger than the NS group, F (2, 56) = 4.03, $p < .02$; $\chi^2 = 33.96$, $p < .03$. There were no other significant diagnostic differences in chronological age, verbal mental age or nonverbal mental age.

Criteria for the participants' expressive language skills for Modules 3 and 4 were identical. Preferred modules for participants between ages 10 and 18 were identified on the basis of the participants' interests in toys such as action figures (available in Module 3) and the ability to tolerate Module 4's less toy-based, more interview-like assessment.

Module 4. As shown in Table III, a group of 45 children and adults, constituting three groups equivalent in verbal IQ, were selected for the validity study for Module 4. All participants in this sample spontaneously used sentences with multiple clauses. They ranged in age from 10 to 40 years. Twenty participants (9 AUT, 7 PDD, 4 NS) were included in reliability analyses. Because most of these participants were adults, verbal IQ was felt to be a better indicator of level of functioning than verbal mental age. There were no significant differences among the groups in chronological age, verbal IQ or performance IQ.

There were no effects of gender, except in Module 3, where there was a main effect of gender on all domain scores. Subsequent gender by diagnosis ANOVAs yielded no effects of gender within diagnosis, suggesting that the findings in Module 3 were due to the disproportionate number of males in the autism and PDD groups and of females in the nonspectrum group, as shown in Table III.

Procedures

For the first 175 referral cases, the ADOS-G was administered as part of a diagnostic assessment by clinical research staff blind to all information except verbal and nonverbal level of functioning. In addition to the research staff, the clinician that had been working with the participant was also present during the administration and scored the protocols. The administrations were videotaped. Two modules (nonoverlapping items were counterbalanced in order) were administered to each participant and scored separately by each of the observers independently. The raters completed coding independently, immediately after the ADOS-G was administered. The module felt to be most appropriate for the participant based on his or her language level was selected after administration.

Of the remaining cases, approximately two thirds were scored live by two examiners as well as videotaped. The remainder were scored live by one examiner only or live by one examiner and from videotape by another. To continue to check reliability, examiners jointly scored approximately one in four administrations. Whether there was joint live or mixed coding depended on where the participant was seen and whether additional data were needed for testing of reliability. Analyses of reliability always included at least one live scoring, with one coder always blind to diagnosis. As discussed in more detail below, sometimes the live scoring was compared to another live scoring and sometimes it was compared to scoring of a videotape.

Twelve different examiners participated in the study. This large number was necessary because data collection took place over several years in several sites, and because of the need to rotate examiners in order to maintain blindness to diagnoses. Prior to their participation in the study, the 12 examiners had observed and coded many live and videotaped ADOS-Gs. Weekly practice coding sessions at the major site were held, in which videotapes were scored and consensus codings for each item reached. Before examiners officially began to collect data, they reached 80% or greater exact agreement with other reliable coders computed item-by-item on three consecutive scorings of Modules 1 or 2 (including one of each module, at least one administration that they carried out, and one administration by another person) and three consecutive scorings of Modules 3 or 4 (with the same restrictions as 1 and 2). In weekly meetings, examiners were consistently able to maintain reliability over 80% exact agreement on item-by-item analyses. The exceptions were two raters (one on-site, one off-site) who showed significant drift in scoring Module 4 protocols and who, when this was identified during consensus scoring, rescored previously administered schedules from videotapes. Examiners from non-Chicago sites had participated in ADOS-G training and achieved reliability on at least two training tapes and the three to eight tapes they provided for this study.

The ADOS-G was usually conducted in a clinic room furnished with several tables and chairs. At least one parent was present during Module 1 and 2 for all younger children and for some participants in Module 3 but not in Module 4. Psychometric testing was conducted first, in most cases, by a different examiner. Only behaviors observed during administration of the ADOS-G protocol were coded. Standard consent procedures, approved by The University of Chicago Institutional Review Board, were followed. Families who participated in the clinic received oral feedback and reports. Participants who were seen only for research received a brief report and a small compensation for expenses.

Reliability and Validity Studies

Overview of Strategy for Item Selection and Algorithm Development. The general strategy for item selection and algorithm development is described below, followed by presentation of the results. Over several preliminary versions of the ADOS-G (including the former ADOS and the PL-ADOS), numerous items were generated and tested. A penultimate draft of each module was then constructed, from which all items in the final version were selected.

Reliability

Reliability of Individual Items

ADOS-G items are typically scored on a 3-point scale from 0 (*no evidence of abnormality related to autism*) to 2 (*definite evidence*). Some items include a code of 3 to indicate abnormalities so severe as to interfere with the observation. For all analyses reported here, scores of 3 were converted to 2. When an item was scored as not applicable (e.g., most often these were language items in Module 1), data were treated as missing

for both validity and reliability analyses. Items that received no more than two scorings other than zero were excluded from reliability analyses. A standard formula for weighted kappas for nonunique pairs of raters was employed (Stata Corp., 1997). Mean weighted kappas (Mk_w greater than .40 were considered to be adequate, with kappas greater than .60 treated as substantial (Landis & Koch, 1977; Stata Corp., 1997). Items for which kappas fell below .40 were excluded, unless an item was felt to be extremely important diagnostically, in which case, comparisons of live scorings only (eliminating comparisons of live and videotape) were checked separately. Codes for these, and other items falling below .50, were rewritten to increase clarification. Rewritten items were then evaluated using at least 20 additional cases from the validity samples.

Interrater reliability was very high for Module 1 items. One item, Behavior When Interrupted, was eliminated from the penultimate draft of Module 1 because of poor reliability. Mean exact agreement of all other items was 91.5%; all items had more than 80% exact agreement across raters. All kappas exceeded .60 (Mk_w = .78), except for items describing repetitive behaviors and sensory abnormalities. These items were less frequently scored as abnormal even within the sample with autism. They also seemed more difficult to score, particularly from video.

Interrater reliability for Module 2 items was also relatively high (mean exact agreement of final item set was 89%). Codes for Social Disinhibition and Language Production and Linked Nonverbal Communication were eliminated because of poor reliability and limited distributions. All other items exceeded 80% exact agreement across rater pairs. Of 26 kappas, 15 exceeded .60 (Mk_w = .70), with the remainder exceeding .50, except for Unusual Repetitive Interests or Stereotyped Behaviors, Unusual Sensory Interest, and Facial Expressions Directed to Others. Kappas for these three items ranged from .46 to .49, with agreements of 81–92%. The coding for facial expression was consequently replaced with a somewhat different version of the same item from Module 3. Codes for other items were edited slightly; reliability checks for at least 20 additional subjects indicated increased agreement for the replaced items.

Items for Module 3 showed similar results for level of agreement as those from Module 2. Items describing Communication of Own Affect, Social Distance, Pedantic Speech (collapsed into the rewritten Stereotyped Speech item) and Emotional Gestures were eliminated because of very poor reliability, even after rewriting. The mean exact agreement was 88.2% across all other items. Of 26 items, 17 received kappas of .60 or better (Mk_w = .65). Three items, Overall Level of Language (k_w = .49), Hand and Finger & Other Complex Mannerisms (k_w = .47), and Compulsions/Rituals (kappa was not computed because of limited distribution), received agreements exceeding 90%, but weighted kappas, in the two cases, of .50. All but four items received more than 80% agreement. Codes for one of these items, Stereotyped Phrases, were rewritten and received adequate reliability in retesting (k_w = .61, % agreement = 92).

In Module 4, items describing Social Disinhibition, Attention to Irrelevant Details, and Pedantic Speech were eliminated because of poor reliability. All other individual items in Module 4 exceeded 80% exact agreement (M = 88.25%). Kappas exceeded .60 for 22 of 31 items (Mk_w = .66), with the remainder exceeding .50 except for Excessive Interest in Highly Specific Topics or Object (k_w = .41) and Responsibility (k_w = .48). These items were retained because of high agreement (85%); codes were rewritten slightly and reassessed.

Overall, interrater item reliability for exact agreement for codes related to social reciprocity was substantial across all modules. Interrater reliability for restricted, repetitive behaviors was consistently adequate but lower than for social items across modules. Item interrater reliability for communication was substantial in Modules 1 and 2, and good, but more variables in Modules 3 and 4, which resulted in the elimination of some items from the original draft. Item interrater reliability for nonspecific abnormal behaviors was substantial across all modules.

Table IV. Intraclass Correlations for Interrater and Test–Retest Reliability

	n	Social	Communication	Social communication	Restricted, repetitive
Inter–rater (all)	97	.93	.84	.92	.82
Live–live	62	.92	.80	.90	.86
Live–video	35	.92	.82	.91	.72
Test–retest	27	.78	.73	.82	.59

Reliability of Domain Scores and Classifications

Intraclass correlations were computed across pairs of raters for algorithm subtotals and total scores for each module separately (the composition of the algorithms is described in the validity sections below) and for the four modules combined. For the social domain, intraclass correlations ranged from .88 to .97 for separate modules. Intraclass correlations for the communication domain ranged from .74 to .90. For the social-communication total used in the algorithm, intraclass correlations ranged from .84 to .98. Intraclass correlations for restricted, repetitive behaviors were somewhat lower, but still high, ranging from .75 to .90. Table IV reports intraclass correlations combined across modules.

Interrater agreement in diagnostic classification for autism versus nonspectrum comparisons based on the ADOS-G algorithm was 100% for Modules 1 and 3, 91% for Module 2, and 90% for Module 4. When PDDNOS participants were included, agreement fell to 93% for Module 1, 87% for Module 2, 81% for Module 3, and 84% for Module 4. Fisher exact tests for comparisons of each diagnosis versus the other two were significant at $p < .01$ in all cases. Disagreements between raters in ADOS-G algorithm diagnoses were almost always between autism and PDDNOS.

Because scoring videotaped observations is often a standard part of research and training with the ADOS-G, interrater reliability for live and videotaped scorings was compared for the 62 participants scored by two raters during live observation to 35 participants scored live and from a video of the same administration. Participants were approximately equally distributed across modules by diagnosis. In the case of the live versus video ratings, the codes made by the examiner immediately after the administration were compared to the scoring of a different researcher watching the same administration on videotape. Analyses were run initially for separate modules to check for any anomalous results and then, because of the relatively small sample and the close comparability across modules in ranges of scores, data were collapsed across modules. As shown in Table IV, there was little difference in the algorithm totals when reliability was computed for two live or live versus video scorings, except in coding of restricted, repetitive behavior. There was no systematic difference in diagnostic classification according to the source (live vs. video) of scoring.

Mean scores for individual items scored live and from video for the same administration for 35 participants were also compared. The mean item difference was less than 0.25 for all items except Imagination/Creativity (in Modules 2 and 4: *M* difference = 0.34 and 0.33, respectively) and Overactivity in Module 2 (*M* difference = 0.28). Scores from live codings for overactivity were lower (less abnormal) by about 0.25 for Modules 1–3 and higher by the same amount for Module 4. Scores from video were lower (less abnormal) than live for Imagination/Creativity (all modules) and lower for video than live for all Stereotyped Behaviors and Restricted Interests codings for Modules 1 and 2. There were no other differences in live versus video scorings of individual items. Differences in domain scores from live versus video scoring averaged 0.50 ($SD = .11$) in the Social domain, 0.28 ($SD = 0.08$) in Communication, and 0.78 ($SD = 0.18$) in the Social-Communication total; these differences were not significant when compared using paired sample *t* tests. Thus, the clinical magnitude of these differences was small and not obviously greater than interrater differences between two live scorings. However, the consistency across Modules 1–3 of subtle biases introduced by the live versus video scoring for Play and Stereotyped Behaviors and Restricted Interests suggests caution when comparing large samples that may systematically vary in the medium from which they were scored.

Test–retest reliability was also assessed for a sample of 27 participants who were administered the same ADOS-G module twice by two different examiners within an average of 9 months. Intraclass correlations are shown in Table IV and indicate excellent stability for Communication and Social domains and the total, with good stability for Stereotyped Behaviors and Restricted Interests. Mean *absolute* differences in domain scores ranged from 1.19 ($SD = 1.6$) in Communication to 1.26 ($SD = 1.39$) in Stereotyped Behaviors and Restricted Interests to 1.78 ($SD = 1.93$) in the social domain to 2.67 ($SD = 1.93$) in the Social-Communication total. Group means changed less than 0.50 in each of the domains except the Social-Communication total ($M = -.94$, $SD = 2.63$). Across modules, Social and Communication domain scores and totals tended to decrease slightly in the second testing, with Stereotyped Behaviors and Restricted Interests scores increasing, though these differences were not significant. Six children's scores (just over 20%) of the sample were associated with changes in ADOS-G diagnoses; three of these changes were associated with general clinical improvements in young children (from autism to ASD) over periods of more than 6 months. Of the three remaining cases in which ADOS-G classification changed in retesting, two children's scores increased, moving them from an ADOS-G classification of ASD to autism and one decreased, moving him from an ADOS-G classification of autism to ASD: all of these cases were children with stable clinical diagnoses of PDDNOS.

Validity Study

Because the goal was to identify a selected number of items to produce an algorithm that operationalized clinical DSM-IV/ICD-10 diagnosis of autism, analyses of validity proceeded in a series of steps.

Validity of Individual Items

First, correlation matrices were generated for all items for each module for each diagnostic group (AUT/MAUT, PDD, NS) separately and together (LAUT for Module 1 was run separately and also combined). Items that were consistently intercorrelated more than .70 for two or more groups in a module that overlapped in conceptualization were targeted for possible elimination. The exception to this rule was Integration of Gaze, which was highly correlated (.76–.93) with several other items and overlapped across diagnostic groups in Module 1. This item was retained in the ADOS-G because of its potential value in future research. It was not included in the algorithm because of redundancy.

Second, exploratory factor analyses were run for each module. One major factor emerged in each module. Almost all social and communication items loaded highly on this factor in each module, accounting for between 52–53 % (Modules 3 and 4) to 72–78 % (Modules 1 and 2) of the variance. Items which did not load primarily on this factor were Response to Joint Attention in Module 2 and Pedantic Speech and Description of Excessive Detail in Modules 3 and 4. Each of these items loaded highly on a factor with verbal mental age or verbal IQ. A second factor, consisting of various combinations of speech (e.g., Stereotyped Speech) and gesture (e.g., Pointing) items accounted for an additional 9–14 % of the variance, though it is important to note that many of these items also loaded higher than .30 on the first factor. Items within the domain of Stereotyped Behaviors and Repetitive Interests tended to load on separate factors that varied considerably across modules. Altogether, these findings were used in the decision to use separate social, communication, and restricted-repetitive sections in the algorithm. Separate analyses of item–total correlations were performed later, once algorithms were finalized. These analyses provide similar information to the factor analyses and are presented in more detail below.

Third, fixed effect analyses of variance (ANOVAs) were performed comparing verbally equivalent samples of autistic and nonspectrum participants. PDDNOS samples were not included in these analyses because it was felt that the initial focus should be on distinguishing the well-established syndrome of autism from nonspectrum disorders. Other than exceptions described below, items that did not yield significant differences were eliminated from the schedule. Items that were retained that did not show group differences included those describing behaviors not expected to be specific to autism, such as Overactivity and Anxiety. These items were retained in order to have a record of behaviors that might affect an observation, without directly contributing to a diagnosis. In addition, Response to Name was retained in Module 1 because it was highly significant as a discriminator of the LAUT group of young, very low-functioning children with autism from the NS group, though not significant for the matched group with autism (MAUT). It was felt that this item was an important communication item for the developmentally younger group and served as a replacement for Stereotyped Speech, which could not be scored in nonverbal children, but was an important item in children with some speech. In addition, two items (i.e., Functional Play, Response to Joint Attention) that differed by diagnosis in Module 1, but not Module 2, were retained in both modules, in order to allow the opportunity to assess progress. These items were excluded from potential diagnostic algorithms in the module in which differences were not significant.

Then, one-way fixed-effect ANOVAs were run comparing the three matched groups (AUT/MAUT, PDDNOS, NS) for each module for each of the retained items, with Scheffé tests used for specific comparisons. When significantly different distributions were indicated by Bartlett's tests, Kruskal-Wallis (1992) tests were run as well; however, in no case, did the additional analyses result in different levels of significance. Except for the nonspecific behavioral items (e.g., Overactivity), the consistent pattern across items for all modules was that scores were highest for the AUT group, lower for PDDNOS and lowest for the NS group. In Modules 1 and 2, 25 to 40% of the items differed significantly across all three groups (AUT, PDDNOS, NS); for Modules 3 and 4, only 10–15 % of the items followed this pattern.

No specific item differed significantly for all three diagnostic groups for all modules, but Unusual Eye Contact and Facial Expressions Directed to Others both differed for all three diagnostic groups for Modules 1, 2, and 3, such that the AUT group scored significantly higher than the PDDNOS group which scored significantly higher than the NS group. Results of specific comparisons for all other items varied across modules, with significant differences occurring more frequently for autism and PDDNOS when compared to the nonspectrum group than when autism and PDDNOS were compared to each other.

Next, lists of items that operationalized each of the criteria in DSM-IV/ICD-10 (except for peer relations) were generated, with items that had yielded significant differences between all possible group combinations identified as highest priority for potential algorithms within each module. Various combinations of items in each domain and each module were considered. Initially, separate lists of items for identifying autism and identifying PDDNOS were proposed; however, in no case was the PDDNOS list more accurate in identifying those clinically diagnosed with PDDNOS than was the autism list when used with adjusted thresholds. Preference was given to items that yielded high levels of discrimination across contiguous modules. On the basis of the factor analyses and these data, algorithms were generated that followed the DSM-IV/ICD-10 strategy of specifying individual totals in the domains of communication and social reciprocity and an overall social-communication total.

Comparison of Domain Scores

Means and standard errors for Social, Communication, Social-Communication totals and Restricted & Repetitive Behaviors domain scores are reported by diagnostic group in Tables V–VIII. ANOVAs and specific comparisons (Scheffé, 1953; Kruskal & Wallis, 1952) comparing distributions for Social domains and Social-Communication totals across diagnostic groups were significantly different for all modules. For Communication and Restricted and Repetitive Behaviors, the AUT and PDDNOS groups generally differed from the nonspectrum group; other specific comparisons were variable.

Item–Total Correlations. Item- "rest" correlations (domain scores minus the particular item) were generated, as well as correlations between domain scores and chronological age, gender, verbal mental age or verbal IQ, and nonverbal mental age or nonverbal IQ. Identical analyses were run for separate groups as well as together. Correlations between potential algorithm items and domains all exceeded .50, with ranges for individual items and the rest of the domain from .62–.88 for the Communication domain (*M*s across modules, .74–.79), and .52–.90 for the Social domain (*M*s across modules, .72–.77). Within the restricted, repetitive domain, item-total correlations exceeded .71 for all items except Unusual Sensory Behaviors in Modules 3 and 4 (.46 and .54, respectively). Social and communication domains were also highly correlated (.82–.89) across modules. Correlations between social-communication totals and restricted, repetitive behavior domain totals were also significant (.51–.60 across modules), but

Table V. Summary Statistics for Module 1 Domain Scores[a]

	Lower autism ($n = 20$)	Matched autism ($n = 20$)	PDDNOS ($n = 17$)	Nonspectrum ($n = 17$)	$F(1, 53)$
Social domain cutoffs (autism = 7; ASD = 4)					
M	11.45_a	10.75_a	8.06_b	1.29_c	91.36
SE	1.47	1.68	2.99	1.61	
Range	9–13	7–14	2–13	0–6	
Communication domain cutoffs (autism = 4; ASD = 2)					
M	7.00_a	5.85_{ab}	4.65_b	1.29_c	42.51
SE	1.30	1.42	1.90	1.21	
Range	5–10	3–8	2–8	0–4	
Social-communication total cutoffs (autism = 12; ASD = 7)					
M	18.45_a	16.60_a	12.71_b	2.59_c	83.14
SE	2.24	2.78	4.59	2.40	
Range	14–23	12–21	4–20	0–9	
Restricted and repetitive domain (no cutoff)					
M	3.50_a	3.05_a	2.53_a	0.53_b	15.87
SE	1.88	1.64	1.55	0.87	
Range	0–6	1–6	0–5	0–3	

[a] When subscripts differ, diagnostic groups are significantly different ($p < .01$) from each other. *F* scores are for univariate comparisons of matched autistic, PDDNOS, and nonspectrum groups. ASD refers to autism spectrum disorder.

Table VI. Summary Statistics for Module 2 Domain Scores[a]

	Lower autism ($n = 21$)	PDDNOS ($n = 18$)	Nonspectrum ($n = 16$)	F (1, 54)
Social domain cutoffs (autism = 6; ASD = 4)				
M	10.76_a	6.61_b	1.81_c	54.02
SE	2.21	2.79	2.83	
Range	6–14	2–12	0–11	
Communication domain cutoffs (autism = 5; ASD = 3)				
M	7.62_a	5.22_{ab}	1.81_c	44.62
SE	1.88	1.66	2.00	
Range	5–10	3–9	0–6	
Social-communication total cutoffs (autism = 12; ASD = 8)				
M	18.38_a	11.83_b	3.63_c	59.02
SE	3.85	3.79	4.69	
Range	11–24	6–19	0–17	
Restricted and repetitive domain (no cutoff)				
*M*ean	2.76_a	1.50_{ab}	0.44_b	13.20
SE	1.58	1.58	0.63	
Range	0–6	0–5	0–2	

[a] When subscripts differ, diagnostic groups are significantly different ($p < .01$) from each other. F scores are for univariate comparisons. ASD refers to autism spectrum disorder.

Table VII. Summary Statistics for Module 3 Domain Scores[a]

	Lower autism ($n = 21$)	PDDNOS ($n = 20$)	Nonspectrum ($n = 18$)	F (1, 58)
Social domain cutoffs (autism = 6; ASD = 4)				
M	9.62_a	7.60_b	1.67_c	85.01
SE	2.29	1.88	1.57	
Range	6–14	4–11	0–5	
Communication domain cutoffs (autism = 3; ASD = 2)				
M	4.90_a	3.65_a	0.61_b	34.51
SE	1.55	2.06	1.14	
Range	3–8	1–8	0–4	
Social-communication total cutoffs (autism = 10; ASD = 7)				
*M*ean	14.52_a	11.25_b	2.28_c	77.44
SE	3.63	3.29	2.22	
Range	10–21	5–16	0–7	
Restricted and repetitive domain (no cutoff)				
M	2.71_a	1.75_{ab}	0.22_b	12.03
SE	2.22	1.41	0.55	
Range	0–8	0–5	0–2	

[a] When subscripts differ, diagnostic groups are significantly different ($p < .01$) from each other. F scores are for univariate comparisons. ASD refers to autism spectrum disorder.

Table VIII. Summary Statistics for Module 4 Domain Scores[a]

	Autism ($n = 16$)	PDDNOS ($n = 14$)	Nonspectrum ($n = 15$)	F (1, 44)
Social domain cutoffs (autism = 6; ASD = 4)				
*M*ean	10.13_a	7.00_b	1.13_c	58.61
SE	2.55	2.60	1.77	
Range	3–14	4–11	0–6	
Communication domain cutoffs (autism = 3; ASD = 2)				
M	5.06_a	3.21_a	0.67_b	29.07
SE	1.91	1.25	1.54	
Range	2–8	1–5	0–6	
Social-communication total cutoffs (autism = 10; ASD = 7)				
M	15.19_a	10.21_b	1.80_c	56.14
SE	3.94	3.42	3.19	
Range	5–22	6–15	0–12	
Restricted and repetitive domain (no cutoff)				
*M*ean	1.94_a	1.07_{ab}	0.20_b	7.41
SE	1.48	1.49	0.56	
Range	0–5	0–5	0–2	

[a] When subscripts differ, diagnostic groups are significantly different ($p < .01$) from each other. F scores are for univariate comparisons. ASD refers to autism spectrum disorder.

lower. Correlations with demographic variables (e.g., age, verbal level) were generally not significant except to the extent that they reflected group differences documented elsewhere. For example, Stereotyped Speech in Module 1 was positively correlated, r (54) = .50, $p <$ .001, with chronological age, because autistic children in Module 1 were older and were more likely to have some language than children with nonspectrum disorders. This language was often stereotyped. No algorithm item correlation with verbal mental age exceeded .30; 37 out of 45 of these correlations were .20 or lower.

Internal consistency was assessed using Cronbach's alpha (Cronbach, 1951). Although the Stereotyped Behaviors and Restricted Interests domain is not included in the algorithm, it was included in these analyses because of the importance of these scores to clinical descriptions of ASD. Cronbach's alphas were consistently highest for the Social domain (.86–.91 for each module), slightly lower for Communication (.74–.84) and lower for Stereotyped Behaviors and Restricted Interests (.63–.65 for Modules 2 and 1; .47–.56 for Modules 4 and 3, respectively), although still indicating good agreement. For the Social-Communication totals, Cronbach's alphas were very high (.91–.94) for all modules.

Finally, Receiver Operating Characteristic (ROC) curves (Siegel, Vukicevic, Elliott, & Kraemer, 1989) were used to provide information concerning where to set cutoffs to indicate different diagnoses for each domain and for each total in each module. Analyses from the ROC curves were used to measure changes in sensitivity and specificity when thresholds for the individual domains or total were raised or lowered. In selecting cutoffs for autism, sensitivity for autism (vs. PDD or NS disorders) and specificity for autism versus comparisons with the nonspectrum group (but not PDD), were given highest priority. In selecting cutoffs for autism spectrum disorder (ASD), sensitivity for diagnosis of either PDDNOS or AUT versus nonspectrum disorders and specificity for PDDNOS versus nonspectrum disorders were considered most important. Thus, a false positive categorization of the scores of a child with a clinical diagnosis of PDDNOS as having autism was considered more acceptable than a false negative categorization of the scores of a child with autism as nonspectrum disorder or the false positive categorization of the scores of a nonspectrum child as having autism or ASD. In general, there were several plausible cutoffs in each module for the differentiation of autism and ASD versus nonspectrum disorders with clear "gaps" between distributions. Conversely, several possible cutoffs in each module for the differentiation between autism and ASD were available in each module because of continuous distributions.

Table IX. Distribution of Participants by ADOS-G Diagnosis and Overall Clinical Diagnosis

	ADOS-G Diagnosis		
Clinical classification	Autism	ASD	Other
	Module 1		
Lower Autism	20	0	0
Autism	19	1	0
PDDNOS	11	5	1
Nonspectrum	0	1	16
	Module 2		
Autism	20	1	0
PDDNOS	8	8	2
Nonspectrum	1	1	14
	Module 3		
Autism	21	0	0
PDDNOS	12	4	4
Nonspectrum	0	1	17
	Module 4		
Autism	13	1	1
PDD-NOS	6	6	2
Nonspectrum	1	0	14

Table X. Sensitivities and Specificities for Different Comparisons Across Modules

	Module 1 ($n = 54$)	Module 2 ($n = 55$)	Module 3 ($n = 59$)	Module 4 ($n = 45$)
	Autism and PDD versus nonspectrum			
Sens.	97	95	90	90
Spec.	94	87	94	93
	Autism versus PDD and nonspectrum			
Sens.	100	95	100	87
Spec.	79	73	68	76
	PDD versus nonspectrum			
Sens.	94	89	80	86
Spec.	94	88	94	93
	Autism versus nonspectrum			
Sens.	100	95	100	93
Spec.	100	94	100	93

Table IX depicts the distribution of participants according to the final ADOS-G algorithm classification and clinical diagnosis and Table X summarizes sensitivities and specificities. In these tables, the term "other" is used for ADOS-G classifications, rather than the designation of nonspectrum disorders used for the clinical diagnoses. This distinction was made in order to reflect our acknowledgment of the limitations of a single observation in the ADOS-G in ruling out evidence from other sources about possible spectrum disorders.

The ADOS-G algorithm yielded the expected classification, for individuals pooled across modules, for nearly 95% of those with autism and 92% of those outside the spectrum, but only categorized 33% of individuals with PDDNOS as having nonautism ASD (with 53% of the PDDNOS sample falling in the range of autism). Total positive and negative predictive values were computed but were not very useful, given the nature of this clinical sample. As shown in Table X, using the algorithms as already defined (i.e., three thresholds: social, communication, social-communication total), the ADOS-G was very effective in discriminating autism from nonspectrum disorders and in discriminating PDDNOS from nonspectrum disorders. However, differentiation of autism and PDDNOS resulted in specificities of .68 to .79. For the discrimination of autism from nonspectrum disorders, when only the social domain or the social-communication total was considered (rather than the three-threshold model), results were very similar to the three-threshold algorithm. However, three thresholds (rather than a simple social or social-communication total) resulted in higher sensitivity for Modules 1 and 3 for autism versus PDD than the simpler approaches, and so were retained.

DISCUSSION

The ADOS-G offers a standardized observation of current social-communicative behavior with excellent interrater reliability, internal consistency and test–retest reliability on the item, domain and classification levels for autism and nonspectrum disorders. Diagnostic validity for autism *versus* nonspectrum disorders, controlling for effects of expressive language level, is also excellent, again documented for individual items, domains, and classification.

An important aspect of interpretation of the ADOS-G is understanding the meaning of cutoffs and algorithms. An individual who meets or exceeds the cutoffs for autism has scored within the range of a high proportion of participants with autism who have similar levels of expressive language in deficits in social behavior and in the use of speech and gesture as part of social interaction (referred to as the ADOS-G communication domain). To meet formal diagnostic criteria for autism, however, an individual must also show evidence of restricted, repetitive behaviors either within the ADOS-G or in another context, and meet criteria for age of manifestation of first symptoms. Thus, there

will be individuals who meet ADOS-G classification criteria for autism who will not receive clinical diagnoses of autism because they do not have restricted, repetitive behaviors or because they have a first manifestation of symptoms after the age of 3 years.

The ADOS-G also introduces an attempt to provide standard thresholds for a broader classification, autism spectrum (the term we have elected to use for the DSM-IV general classification of pervasive developmental disorders), when cutoffs for autism are not met. A participant who meets ADOS-G criteria for ASD shows significant abnormalities in social reciprocity and the use of gesture and language in social interaction. Evidence from genetic (Bailey *et al.*, 1995) and longitudinal studies, as well as the data presented here, suggests that autism and other pervasive developmental disorders are on a continuum of severity, with little evidence of qualitative differences between the categories, especially if language level is considered separately. At the outset of this research, it was anticipated that distinct patterns of deficits for autism and at least some participants with PDDNOS could be identified. However, abnormalities that defined PDDNOS were consistently similar in quality to those of persons with diagnoses of autism in each of the four modules. In fact, more individuals with clinical diagnoses of PDDNOS received an ADOS-G classification of autism than of the broader category of autism spectrum disorder. Although the sizes of each sample within a module are relatively small, the fact that distributions between autism and PDDNOS were continuous and overlapping for all items and algorithm domain scores, supports a conceptualization of a spectrum of autistic disorders, rather than discrete differentiations between a narrow autism classification and PDDNOS as a separate entity.

Unlike parent-report measures such as the ADI-R (Lord *et al.*, 1994) and measures that can be completed retrospectively, such as the CARS (Schopler, Reichler, DeVellis, & Daly, 1980), the ADOS-G only provides a measure of current functioning. Thus, individuals may have met criteria for autism at younger ages, but fail to meet current ADOS-G criteria. For diagnostic systems that emphasize lifetime criteria, these individuals would still be considered to have autism. The ADOS-G provides a reliable way of differentiating current performance of high- and low-scoring individuals. On a group level, autism and PDDNOS diagnoses were consistently differentiated quantitatively by the ADOS-G domain scores. However, at an individual level, there was significant overlap. Thus, more information and additional approaches are needed to address the question of whether, and if so, how differentiation between a narrower definition of autism and a broader conceptualization, autism spectrum/PDD, including Asperger syndrome, can be made on the basis of a relatively brief observation.

On the basis of the present data, it was decided to require individuals to meet cutoffs in both social and communication domains to receive ADOS-G algorithm classifications of autism spectrum disorder. Thus, an individual could meet DSM IV/ICD-10 criteria for PDDNOS by having significant abnormalities in social reciprocity (as evidenced by a high score in the social domain) and having restricted, repetitive behaviors (as evidenced during the ADOS-G or through other methods), and not meet ADOS-G criteria because of failure to meet the ADOS-G communication cutoff. It is important to recognize that high scores in either the social or communication domain indicate clinically significant difficulties. Such scores require further clinical consideration, even if autism or autism spectrum disorder cutoffs are not met.

Nonalgorithm items were retained in the instrument for a number of reasons. First, the standardization and reliability statistics were based on scoring all items. From past experience, it seems likely that scoring of some items affects scoring of others because of the emphasis on not scoring the same behavior twice. Second, some of the nonalgorithm items were deliberately retained in order to allow comparability across modules and measurement of change, even when a behavior is not diagnostically significant for the ADOS-G (e.g., Shared Enjoyment in the later modules; repetitive behaviors). Third, some of the nonalgorithm items are codings of behaviors that are not specific to autism (e.g., Anxiety) but that are important clinical observations. Four, future research may result in changing conceptualizations of the critical features of autism beyond those that were used to generate the present algorithms. Information about homogeneous samples of children with other disorders, such as developmental language disorder or nonverbal learning disabilities or Williams syndrome, will also clarify the usefulness of the ADOS-G items and algorithm domains in differentiating autism-specific disorders.

Interpretation of ADOS-G scores is based on the assumption that a valid sample of behavior is collected, that a similar sample of behavior would be elicited by another examiner and at another time, and that the examiner can code this behavior in a fashion similar to other examiners using the same codes. In most cases, parents/caregivers will have participated in the administration of Modules 1 and 2 and so can provide information as to how typically their child behaved. Watching and participating in an ADOS-G can also be helpful for parents in understanding the basis of their child's

diagnosis. For older children and adults, depending on the participant's wishes, it may be helpful to have a parent/caregiver or another person who knows the individual well observe the ADOS-G from another room or on video. In any case, the examiner needs to judge whether factors extraneous to the social demands of the ADOS-G may have influenced the assessment.

In addition, the examiner will want to consider if there are aspects of the participant's behavior that may have affected his or her scores, even when the ADOS-G seemed to provide a valid sample of behavior. Specific effects of cultural factors have not yet been addressed systematically in research, though the ADOS-G has been used in many European and some Asian countries. For valid scoring, the examiner should consider the appropriateness of a child or adult's behavior within that individual's cultural context.

The goal of the ADOS-G is to provide standardized contexts in which to observe the social-communicative behaviors of individuals across the life-span in order to aid in the diagnosis of autism and other pervasive developmental disorders. For this reason, the ADOS-G domain or total scores may not be an ideal measure of response to treatment or of developmental gains, especially in the later modules. However, on an individual level, there are several strategies that clinicians or researchers may take to measure how behaviors may have changed over time. If an individual has been administered the same module more than once, raw scores on individual items and on algorithm domains can be compared. If an individual has changed modules, comparison of raw domain scores is not meaningful. However, scores on items that remain constant across modules (about two thirds of each contiguous module; see Table II) can be compared. Behavioral changes may also be indicated by changes in codes not specific to autism, such as Overactivity and Anxiety. In addition, more detailed coding of communication samples or particular behaviors (e.g., pragmatics, sentence structure, gestures) may also be carried out from videotapes of the ADOS-G. Other observational coding schemes that address specific aspects of behavior in more detail may also be applied using the ADOS-G as a way of obtaining a discrete sample of behavior in standard contexts.

Often, clinicians carrying out diagnostic assessments may wish to make programming suggestions for parents/caregivers, therapists, or teachers. Many of the activities and codes of the earlier modules have fairly straightforward implications both for how to teach an individual child and for the content of appropriate goals. For example, Module 1 provides opportunities for children to make requests in a number of circumstances, including requests for action (i.e., the examiner to blow a balloon), requests for food, requests to continue a social game, and requests for an object or activation of that object (i.e., operating a bubble gun). Noting how children make requests and in what circumstances they are most easily able to communicate their interest or needs, allows the clinician to create goals to teach new request behaviors and to helping the children generalize existing behaviors across contexts.

Generating programming goals from Modules 3 and 4 may be somewhat more complex, because fewer codes describe specific behaviors that may be usefully taught in a direct fashion. Realizing the degree to which adults with autism have limited insight into the nature of social relationships, or having the opportunity to observe adolescents describing the emotions of the main characters in a story, can be helpful in representing the strengths they may have and difficulties they experience in social interaction.

Overall, the ADOS-G offers a more comprehensive opportunity for standardized observation than previously available. Replication of psychometric data with additional samples including more homogeneous nonautistic populations and more individuals with pervasive developmental disorders who do not meet autism criteria, establishing concurrent validity with other instruments, evaluation of whether treatment effects can be measured adequately, and determining its usefulness for clinicians are all pieces of information that will add to our understanding of its most appropriate use.

ACKNOWLEDGMENTS

We acknowledge the help of the children and adults and their families who participated in the research and in the training workshops. The comments and suggestions of Lennart Pedersen in particular and many other colleagues who led and attended ADOS-G workshops in the U.S., Sweden, Denmark, and the U.K., and the International Consortium on the Genetics of Autism meetings are much appreciated. Contributions of cases by Ami Klin at Yale, Ann Le Couteur at Newcastle, and Rachel Yeung-Courchesne and Senia Pizzo at UCSD were very helpful. The technical assistance of Terri Rossi, Kathleen Kennedy Martin, Sippi Katz-Janssen, Nichole Felix, Sam Park, and Nishchay Maskay in production of the manual and protocols is also gratefully acknowledged. Data collection by Cynthia Brouillard, Steve Guter, Amy Jersild, Jane Nofer, Cory Shulman, Elisa Steele, Audrey Thurm, Saritha Mathew Vermeer, and Marrea Winnega was also critical to the evaluation of the instrument.

REFERENCES

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author

Bailey, A., Le Couteur, A., Gottesman, I., Bolton, P., Simonoff, E., Yuzda, E., & Rutter, M. (1995). Autism as a strongly genetic disorder: Evidence from a British twin Study. *Psychological Medicine, 25,* 63–77.

Berument, S. K., Rutter, M., Lord, C., Pickles, A., Bailey, A., MRC Child Psychiatry Unit, & Social, Genetic and Developmental Psychiatry Research Centre (2000). *Autism Screening Questionnaire: Diagnostic validity. British Journal of Psychiatry, 175,* 444–451.

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

DiLavore, P., Lord, C., & Rutter, M. (1995). Pre-Linguistic Autism Diagnostic Observation Schedule (PL-ADOS). *Journal of Austism and Developmental Disorders, 25,* 355–379.

Dunn, L. M., & Dunn, L. M. (1981). *Peabody Picture Vocabulary Test (Rev.).* Circle Pines, MN: American Guidance Service.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test (3rd ed.).* Circle Pines, MN: American Guidance Service.

Elliott, C. D. (1990). *Differential Abilities Scale (DAS).* San Antonio, TX: Psychological Corp.

Happé, F. G. E. (1995). The role of age and verbal ability in the theory of mind task performance of subjects with autism. *Child Development, 66,* 843–855.

Kobayashi, R., Murata, T., & Yoshinaga, K. (1992). A follow-up study of 201 children with autism in Kyushu and Yamaguchi areas, Japan. *Journal of Autism and Developmental Disorders, 22,* 395–411.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association, 47,* 583–621.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174.

Le Couteur, A., Rutter, M., Lord, C., Rios, P., Robertson, S., Holdgrafer, M., & McLennan, J. D. (1989). Autism Diagnostic Interview: A semistructured interview for parents and caregivers of autistic persons. *Journal of Autism and Developmental Disorders, 19,* 363–387.

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

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

Lord, C., Rutter, M., & Le Couteur, A. (1994). Autism Diagnostic Interview-Revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders, 24,* 659–685.

Mahoney, W., Szatmari, P., Maclean, J., Bryson, S., Bartolucci G., Walter, S., Hoult, L., & Jones, M. (1998). Reliability and accuracy of differentiating pervasive developmental disorder subtypes. *Journal of the American Academy of Child and Adolescent Psychiatry, 37,* 278–285.

Mesibov, G. B., Schopler, E., & Hearsey, K. A. (1994). Structured teaching. In E. Schopler & G. B. Mesibov (Eds.), *Behavioral issues in autism: Current issues in autism.* New York: Plenum Press.

Mullen, E. M. (1995). Mullen Scales of Early Learning: AGS edition. Circle Pines, MN: American Guidance Service.

Murray, H. A. (1938). *Explorations in personality.* New York: Oxford.

Raven, J. C. (1956). *Guide to using the Coloured Progressive Matrices.* London: H. K. Lewis.

Raven, J. C. (1960). *Guide to using the Standard Progressive Matrices.* London: Lewis.

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika, 40,* 87–104.

Schopler, E., Reichler, R., DeVellis, R., & Daly, K. (1980). Toward objective classification of childhood autism: Childhood Autism Rating Scale (CARS). *Journal of Autism and Developmental Disorders, 10,* 91–103.

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.

Stata Corporation (1997). *Stata statistical software: Release 5.0.* College Station, TX: Stata Corp.

Venter, A., Lord, C., & Schopler, E. (1992). A follow-up study of high-functioning autistic children. *Journal of Child Psychology and Psychiatry, 33,* 489–507.

Wechsler, D. (1984). *Manual for the Wechsler Adult Intelligence Scale-Revised.* San Antonio, TX: Psychological Corp.

Wechsler, D. (1991). *Manual for the Wechsler Intelligence Scale for Children-Third Edition.* San Antonio, TX: Psychological Corp.

World Health Organization. (1993). *The International Classification of Diseases-10th Revision: Classification of mental and behavioral disorders.* Geneva: Author.