0.55 for individuals younger than 5 years. Specificity was consistently high regardless of age and intelligence. ADI-R-JV was shown to be a reliable tool, and has sufficient discriminant validity and satisfactory diagnostic validity for correctly diagnosing AD, although the diagnostic validity appeared to be compromised with respect to the diagnosis of younger individuals.

Keywords Autism · ADI-R · Reliability · Validity · Japan

Introduction

Autistic disorder (AD) is defined by irregularities in three behavioral domains, namely, deficits in reciprocal social interaction, deficits in communication, and restricted and repetitive behaviors and interests (American Psychiatric Association 2000). AD is classified as an autism spectrum disorder (ASD), an umbrella term that encompasses AD and pervasive developmental disorder not otherwise specified (PDDNOS). The reported prevalence estimates of AD or ASD have been increasing (Fombonne 2009; Williams et al. 2006), with the prevalence of ASD now thought to be between 1 and 2 per 100 school children in the United Kingdom (Baron-Cohen et al. 2009) and in Japan (Kawamura et al. 2008), and even higher in South Korea (Kim et al. 2011). The observed change in prevalence estimates has been suggested to be an artifact due to increased awareness of ASDs, changes in diagnostic precision, and recent trends toward earlier diagnosis (Kočovská et al. 2012; Parner et al. 2008; Waterhouse 2008). Such observations have hastened the worldwide demand for reliable and valid methods of identifying ASD.

A number of questionnaires, interviews, and observation schedules have been developed to assist clinicians and researchers in the diagnostic assessment of specific behaviors found in individuals with AD or ASD. Among these instruments, Autism Diagnostic Interview-Revised (ADI-R (Lord et al. 1994)) is a structured, investigator-based interview directed to caregivers for the detection of AD in a research context. ADI-R has been widely used, and its reliability and validity have been examined in the original as well as in non-English versions (Cicchetti et al. 2008; Hill et al. 2001; Lampi et al. 2010; Lord et al. 1994; Mildenberger et al. 2001).

Discussions of ADI-R have accumulated, particularly as regards its diagnostic validity. Despite the fact that ADI-R provides a good to excellent level of sensitivity for diagnosing and predicting AD among varying samples (de Bildt et al. 2004; Gray et al. 2008; Lampi et al. 2010; Lord et al. 1994, 2006; Tomanik et al. 2007), studies have pointed out compromised diagnostic validity in certain types of examinees, such as younger children, because some symptoms are not evident at an early age (Cox et al. 1999; Rutter et al. 2003). This observation is of particular relevance among individuals with ASD other than AD (Gilchrist et al. 2001). On the other hand, as the algorithm-based diagnosis with ADI-R is made with reference to current as well as past behaviors, caregivers of examinees tend to report fewer symptoms when examinees are in adolescence or early adulthood (McGovern and Sigman 2005). Furthermore, depending on the level of function, ADI-R diagnoses of AD among children exhibiting a cognitive delay are less likely to conform to clinical or other types of research-related diagnosis (de Bildt et al. 2004), such as those based on Autism Diagnostic Observation Schedule (ADOS (Lord et al. 2000)). It should be noted that the use of ADOS alone has limited predictability (Lord et al. 2006). Considering these pitfalls, some groups have recommended that not a single source but rather multiple sources of information, including both ADI-R and ADOS, should be consulted when establishing a diagnosis of ASD or AD (Le Couteur et al. 2008; Lord et al. 2006), particularly in a research context. It follows

K. Ogasahara
Comprehensive Educational Science Division, Tokyo Gakugei University, Tokyo, Japan

T. Miyachi
Department of Pediatrics, Nagoya City University Hospital, Nagoya, Japan

I. Tani
Faculty of Humanities, Tokai Gakuen University, Aichi, Japan

M. Inoue · K. Nomura
Graduate School of Medical Science, Center for Clinical Psychology, Tottori University Faculty of Medicine, Tottori, Japan

K. Nomura
Department of Child Psychiatry, Hamamatsu University School of Medicine, Handayama 1 Higashiku, Hamamatsu 431-3192, Japan

T. Hagiwara
Division of Special Education, Department of Education and Development, Hokkaido University of Education, Asahikawa, Japan

T. Uchiyama
Faculty of Human Development and Culture, Fukushima University, Fukushima, Japan

H. Ichikawa
Tokyo Metropolitan Children's Medical Center, Tokyo, Japan

S. Kobayashi · K. Miyamoto
Department of Pediatrics, Kosai City Hospital, Kosai, Japan

N. Takei
King's College of London, Institute of Psychiatry, London, UK

that the foundation of reliability and validity of ADI-R is important in countries such as Japan, where such diagnostic tools have not been readily available.

ADI-R in particular was unavailable in Japan until 2005, when the present authors translated the WPS Edition of ADI-R (Rutter et al. 2003) into Japanese, at which time the back-translation was confirmed to be congruent with the original version by the developers of ADI-R. However, the reliability and validity of the Japanese version had remained unexamined to date.

Therefore, in the present study, the authors aimed to test the inter-rater reliability and discriminant and diagnostic validity of ADI-R, Japanese Version (ADI-R-JV). The inter-rater reliability was assessed using two types of agreement measures: the weighted Kappa (Kw) and intra-class correlation coefficient (ICC) of diagnostic algorithm item scores of two independent interviewers. Discriminant validity was assessed by comparing mean scores of diagnostic algorithm items/subdomains/domains between individuals with and without a consensus clinical diagnosis. Diagnostic validity in this study refers to agreement between the algorithm diagnosis based on ADI-R-JV and a consensus clinical diagnosis. The sensitivity, specificity, positive predictive value, and negative predictive value were calculated to assess this agreement.

For our assessment, we hypothesized the following.

1. Good to excellent inter-rater reliability in terms of the Kw and ICC of ADI-R-JV would be observed, which would be consistent with the published literature (Cicchetti et al. 2008; Hill et al. 2001; Lord et al. 1994).
2. The discriminant validity of ADI-R-JV would be sufficient, with higher mean scores of diagnostic algorithm items among individuals with AD than among those without AD (Lampi et al. 2010; Lord et al. 1994). That is, it was expected that AD scores > non-ASD scores, and AD scores > PDDNOS scores.
3. The diagnostic validity of ADI-R-JV would be satisfactory yet compromised among younger individuals and individuals with intellectual disabilities (Cox et al. 1999; de Bildt et al. 2004; Rutter et al. 2003).

Methods

Participants and Diagnostic Procedure

Reliability Study

To enroll study subjects, we recruited participants from 3 research sites, namely, 2 developmental, university-affiliated clinics and 1 research center. Basically, these clinics are open for referrals from local health practitioners. Participants were selected on the basis of the cumulative number of participants thus far enrolled (targeted N = 30), age (kindergarteners or school-age children/adolescents under 20 years of age), clinical diagnosis (confirmed or suspected diagnosis of ASD), and the provision of consent to participate in the study voluntarily, including videotaping. Thus, purposive sampling was incorporated into the study design.

For the reliability study, we recruited 35 individuals who were referred to one of our research sites between December 1, 2006 and November 30, 2010 (Table 1). Among them, 31 individuals had been already suspected of having ASD by their local health practitioners and had been referred to our institutions for a more definitive diagnosis. Soon after participating in this study, these participants underwent a clinical assessment based on DSM-IV-TR (American Psychiatric Association 2000) assessment, conducted by one of the authors. After the detailed clinical assessments were complete and comprehensive caregiver interviews were conducted in order to collect the developmental history of the participants, our research team provided consensus clinical diagnoses based on DSM-IV-TR. Our research team included clinical experts with more than 3 years of experience in pediatrics or in child neurodevelopmental practices and in assessing individuals with ASD (5 certified clinical psychologists, 3 child psychiatrists, and 4 pediatricians were involved). A total of 31 individuals were confirmed to have a consensus clinical diagnosis of ASD, namely, AD (N = 12) or PDDNOS (N = 19). The remaining 4 individuals were referred to our research sites on the basis of suspected intellectual impairment, and they were confirmed not to have a diagnosis of ASD according to the same diagnostic procedures as those used for the confirmed ASD cases.

The 35 clinically referred individuals were also examined with respect to cognitive measures. For those subjects who were age 5 or older, the Japanese version of the Wechsler Intelligence Scale for Children, third edition (WISC-III: (Wechsler et al. 1992)) or the Tanaka-Binet intelligence scale (Tanaka Institute of Education 1987) was used to estimate the intelligence quotient (IQ). For individuals younger than 5 years old, a standardized developmental test, the Kyoto Scale of Psychological Development (Koyama et al. 2009), was adopted to estimate development quotient (DQ). Among the 31 individuals with ASD, 6 had a full-scale IQ/DQ of lower than 70. Among the 4 non-ASD clinical individuals, all had a full-scale IQ/DQ of lower than 70.

In addition to the clinically referred individuals, 16 kindergarteners and school-age children exhibiting typical development were also invited to participate in the study as

Table 1 Reliability study: characteristics of the sample studied

	Clinically referred individuals [N = 35]	Control individuals [N = 16]	Statistics
Age in years			
Range	3–18	3–14	
Median	5.0	5.0	
Mean (SD)	8.7 (5.2)	7.0 (3.8)	$t(49) = 1.16$, $p = 0.25$
Gender (F:M)	5:30	4:12	Chi-square(1) = 0.84, $p = 0.36$
Full scale IQ/DQ[a]			
Number of individuals with cognitive delay (IQ/DQ < 70)	10 (29 %)	0 (0 %)	Chi-square(1) = 5.67, exact $p = 0.02$
Range	42–118	86–124	
Median	81	102.5	
Mean (SD)	81.9 (22.6)	102.0 (11.6)	$t(44) = 2.85$, $p < 0.001$
DSM-IV-TR diagnosis			
Autistic disorder	11 (31 %)	0	
Autistic disorder + mental retardation	1 (3 %)	0	
Pervasive developmental disorder, not otherwise specified	14 (20 %)	0	
Pervasive developmental disorder, not otherwise specified + Mental retardation	5 (14 %)	0	
Mental retardation	4 (11 %)	0	
Major depressive disorder	0	1 (6 %)	
Adjustment disorder	0	1 (6 %)	
No psychiatric diagnosis	0	14 (88 %)	
ADI-R score (based on data derived from a first examiner)			
Domain A			
Range	5–28	0–7	
Median	18	3.5	
Mean (SD)	15.9 (6.6)	3.3 (2.8)	$t(49) = 7.16$, $p < 0.001$
Domain BV[a]	[N = 23]	[N = 14]	
Range	3–14	0–8	
Median	7	2	
Mean (SD)	7.3 (3.6)	3.3 (2.9)	$t(35) = 6.94$, $p < 0.001$
Domain BNV[b]	[N = 12]	[N = 2]	
Range	1–12	0–1	
Median	8	0.5	
Mean (SD)	6.9 (4.5)	0.5 (0.7)	$t(12) = 1.96$, $p = 0.07$
Domain C			
Range	0–11	0–4	
Median	3	0.5	
Mean (SD)	3.5 (2.5)	1.3 (1.5)	$t(49) = 3.35$, $p = 0.002$

[a] 5 Individuals, all aged 6 years or older, in the control individuals have no data on IQ/DQ. The school records of these participants were carefully checked and we regarded their histories as equivalent to a lack of cognitive delay

[b] Verbal subjects (defined as a score of 0 on item 30 "overall level of language")

[c] Non-verbal subjects (defined as a score of 1 or 2 on item 30)

control individuals. The control groups was recruited via a notice published in newspapers local to three of our research sites, where the clinically referred individuals for the reliability study had also been enrolled. The characteristics of these control individuals are given in Table 1. Considering the male predominance among clinically

referred children, boys were intentionally oversampled. The control subjects underwent clinical assessment based on DSM-IV-TR in an interview conducted by one of the authors, and the results were later confirmed by our research team according to the same procedures as those described above. Among the control subjects, 1 individual had a diagnosis of major depressive disorder, and 1 had a diagnosis of adjustment disorder. All 16 control individuals were also examined either using WISC-III, the Kyoto Scale of Psychological Development, or the Tanaka-Binet intelligence scale, depending on the subject's mental age, and none of the control subjects were confirmed to have any cognitive delays.

In sum, the enrolled participants comprised two groups (Table 1): 35 clinically referred individuals and 16 control individuals. The mean age of these two groups did not differ significantly (8.7 [SD 5.2] vs. 7.0 [SD 3.8]; $t(49) = 1.15$, $p = 0.25$), and the F:M ratio did not differ (F:M = 5:30 vs. 4:12; Chi-square (1) = 0.84, $p = 0.35$), although the mean IQ/DQ differed significantly (81.9 [SD 22.6] vs. 102.0 [SD 11.6]; $t(44) = 4.9$, $p < 0.001$).

Validity Study

To collect a sufficient number of clinically referred individuals in this sub-study, 6 additional research sites were involved (4 developmental, university-affiliated clinics, 1 pediatric clinic at a general hospital, and 1 privately run clinic for child psychiatry), together with the three research sites also involved in the reliability study. The mode of purposive selection of study participants was the same as that adopted in the reliability study except that in the validity study, the targeted number of participants was larger (N = 200), and the recruitment period was longer (September 1, 2006 and March 31, 2011). To capture any differences between the two recruitment methods used for the two sub-studies, we compared 35 clinically referred individuals enrolled in the reliability study and an additional 200 clinically referred individuals (not shown in the Table). This comparison did not reveal any significant difference in the F:M ratio (F:M = 5:30 vs. 42:158; Chi-Square(1) = 0.84, $p = 0.36$), no significant difference in mean age (mean = 8.7 (SD 5.2) vs. 10.5 (SD 4.9) years; $t(233) = 0.61$, $p = 0.54$), and no significant difference in mean DQ/IQ (81.9 (SD 22.6) versus 89.2 (SD 24.8); $t(233) = 1.62$, $p = 0.11$) between the two groups of individuals. Therefore, we regarded these two groups as basically the same in terms of background characteristics. We then combined the two groups and considered them as feasible for the analysis. A total of 235 clinically referred individuals were enrolled in the validity study.

To establish the group of control individuals, 66 kindergarteners and school-age children exhibiting typical development were also invited to participate in this study. Participants were recruited through a notice placed in local newspapers that serve the regions of the nine research sites at which the 235 clinically referred individuals were also enrolled. As a group, these individuals were identical in terms of mean age, F:M ratio, and mean IQ/DQ to the 16 control individuals enrolled in the reliability study, and as such, they were combined as a single control group of individuals. As a result, for the validity study, we investigated 235 clinically referred individuals and 82 control individuals (Appendix Table 2 in supplementary materials). The mean age of the 235 clinically referred individuals was older than that of the 82 control individuals (10.3 (SD 4.9) vs. 6.5 (SD 3.8) years; $t(315) = 6.42$, $p < 0.001$), and the mean full-scale IQ/DQ of the clinically referred individuals (86.6 (SD 23.0) vs. 100.2 (SD 13.3); $t(310) = 4.65$, $p < 0.001$) was lower than that of the control individuals. There were significantly more male individuals among the clinically referred individuals than among the control individuals (F:M = 47:188 vs. 34:48; Chi-Square(1) = 14.7, $p < 0.001$; see Appendix Table 2 in supplementary materials).

As was done in the reliability study, 235 clinically referred individuals and 82 control individuals underwent a clinical assessment based on DSM-IV-TR (American Psychiatric Association 2000) conducted by one of the authors, and diagnoses, if any, were confirmed by our research team and were established as a DSM-IV-TR-based consensus clinical diagnosis. Among the 235 clinically referred individuals, 227 were confirmed to have ASD, namely, AD (N = 138) or PDDNOS (N = 89) as the consensus clinical diagnoses. The remaining 8 individuals were assessed as not having ASD. Among the 82 control individuals, none had a diagnosis of ASD; however, 1 had a diagnosis of major depressive disorder, 1 had social phobia, 1 had attention deficit/hyperactive disorder not otherwise specified, and 1 had adjustment disorder. To measure IQ/DQ, WISC-III, Tanaka-Binet intelligence scale, or Kyoto Scale of Psychological Development was employed. Among the 82 control individuals, 12 had no IQ/DQ records; the school records of these participants were carefully checked and we regarded their histories as equivalent to a lack of cognitive delay.

Finally, the 235 clinically referred individuals and 82 control individuals were combined and re-grouped into the three following diagnostic groups based on a consensus clinical diagnosis (Table 2): 138 individuals with AD, 89 with PDDNOS, and 90 with non-ASD. Group comparisons of mean age across the three groups revealed a significantly higher value in the AD group than in the other two groups (AD 11.7 [SD 4.3], PDDNOS 8.5 [SD 5.1], non-ASD 6.4 [SD 3.7]; $F(2, 314) = 42.1$, $p < 0.001$). Likewise, the F:M ratio of the three groups showed a significant difference

Table 2 Validity study: characteristics of the sample studied

	(1) AD [N = 138]	(2) PDDNOS [N = 89]	(3) Non-ASD [N = 90]	Statistics
Age in years				
Range	2–19	2–19	2–17	
Median	11.8	8.0	5.0	
Mean (SD)	11.7 (4.3)	8.5 (5.1)	6.4 (3.7)	$F(2, 314) = 42.9$, $p < 0.001$ $1 > 3$: $p < 0.001$ $2 > 3$: $p < 0.001$ $1 > 2$: $p < 0.001$
Gender (F:M)	18:120	25:64	38:52	Chi-square(2) = 24.8, $p < 0.001$
Number of individuals with cognitive delay (IQ/DQ < 70)	18 (13 %)	9 (10 %)	8 (9 %)	Chi-Square(2) = 1.1, $p = 0.59$
DSM-IV-TR diagnosis				
Autistic disorder	120 (87 %)	0	0	
Autistic disorder + mental retardation	18 (13 %)	0	0	
Pervasive developmental disorder, not otherwise specified	0	80 (90 %)	0	
Pervasive developmental disorder not otherwise specified + mental retardation	0	9 (10 %)	0	
Mental retardation	0	0	8 (9 %)	
Major depressive disorder	0	0	1 (1 %)	
Social phobia	0	0	1 (1 %)	
Attention deficit/hyperactive disorder, not otherwise specified	0	0	1 (1 %)	
Adjustment disorder	0	0	1 (1 %)	
No psychiatric diagnosis	0	0	78 (87 %)	
Full scale IQ/DQ[a]				
Range	41–140	42–131	45–132	
Median	87.5	90	93	
Mean (SD)	88.4 (22.8)	87.9 (20.7)	90.8 (23.1)	$F(2, 302) = 0.2$, $p = 0.82$
ADI-R score				
Domain A				
Range	8–30	3–28	0–11	
Median	20	13	1	
Mean (SD)	19.9 (5.3)	14.8 (6.4)	2.3 (2.7)	$F(2, 314) = 330.6$, $p < 0.001$ $1 > 3$: $p < 0.001$ $2 > 3$: $p < 0.001$ $1 > 2$: $p < 0.001$
Domain BV[b]	[N = 116]	[N = 68]	[N = 79]	
Range	3–25	2–21	0–12	
Median	14	8.5	1	
Mean (SD)	14.3 (4.1)	9.7 (4.4)	2.5 (3.2)	$F(2, 260) = 210.9$, $p < 0.001$ $1 > 3$: $p < 0.001$ $2 > 3$: $p < 0.001$ $1 > 2$: $p < 0.001$
Domain BNV[c]	[N = 22]	[N = 21]	[N = 11]	
Range	0–14	1–12	0–9	
Median	10	6	1	

Table 2 continued

	(1) AD [N = 138]	(2) PDDNOS [N = 89]	(3) Non-ASD [N = 90]	Statistics
Mean (SD)	12.6 (4.9)	9.0 (4.4)	2.3 (2.5)	$F(2, 51) = 21.0$, $p < 0.001$ $1 > 3$: $p < 0.001$ $2 > 3$: $p < 0.005$ $1 > 2$: $p = 0.02$
Domain C				
Range	0–12	0–12	0–9	
Median	5	2	0	
Mean (SD)	5.5 (2.4)	2.9 (2.5)	1.1 (1.8)	$F(2, 314) = 106.6$, $p < 0.001$ $1 > 3$: $p < 0.001$ $2 > 3$: $p < 0.001$ $1 > 2$: $p < 0.001$

NS not significant

[a] 12 individuals, all aged 6 years or older, in the Non-ASD group have no data on IQ/DQ. The school records of these participants were carefully checked and we regarded their histories as equivalent to a lack of cognitive delay

[b] Verbal subjects (defined as a score of 0 on item 30 "overall level of language")

[c] Non-verbal subjects (defined as a score of 1 or 2 on item 30)

(AD 18:120, PDDNOS 25:64, Non-ASD 38:52; Chi-Square(2) = 24.8, $p < 0.001$). The mean IQ/DQ did not differ across the three groups (AD 90.8 [SD 23.0], PDD-NOS 87.9 [SD 20.1], Non-ASD 88.3 [SD 88.3]; $F(2, 302) = 0.2$, $p = 0.82$), and the proportion of individuals with an IQ/DQ of less than 70 did not show any statistically significant departures from the expected values (AD 13 %, PDDNOS 10 %, Non-ASD 9 %, Chi-Square(2) = 1.07, $p = 0.59$).

With ADI-R-JV, an algorithm diagnosis of AD was provided if the sum scores of all of four domains (A, B, C, and D) met the criteria (equal to or exceeding the cutoff for each domain) as described in the original guidelines (Rutter et al. 2003).

Interviews Using ADI-R-JV

All caregivers of participants in this study were interviewed using ADI-R-JV within a 2-month period after the participants had taken part in the study. These interviews were conducted either by one of the present authors (KJT, KM, AY, SS) who established the research reliability of the original ADI-R together with the developers based on intensive training sessions at the training sites, namely, the interviewers reached more than 90 % exact agreement with the ADI-R trainers (Risi et al. 2006), or by the authors who were supervised by the authors KJT, KM, AY, or SS when the interview using ADI-R-JV was conducted. In this study, the same standard of agreement was achieved across all members of the research team who conducted ADI-R-JV. In total, 8 of the present authors were entitled to conduct interviews using ADI-R-JV, and thus were regarded as ADI-R-JV interviewers for the current study.

For the reliability study, all ADI-R-JV interviews were first conducted by one of four interviewers (KJT, KM, AY, SS), and all interviews were videotaped. Each tape was assessed independently by another rater from the same group of four interviewers, and all combinations of the four raters were equally likely. For the validity study, only one out of 8 interviewers conducted an ADI-R-JV interview, and that interviewer was blind to the consensus clinical diagnosis of the examinee. All 8 interviewers assessed participants at each research site on a random basis.

Analyses

Construction of ADI-R-JV Diagnostic Algorithm

ADI-R diagnostic algorithm consists of the following 4 domains: (A) Qualitative abnormalities in reciprocal social interaction; (B) Qualitative abnormalities in communication; (C) Restricted, repetitive, stereotyped patterns of behavior; and (D) Abnormality of development evident at or before 36 months. Domains A, B, and C correspond to the three groups of symptoms described in the DSM-IV-TR (American Psychiatric Association 2000). Domain A

consists of 4 subdomains covering 16 algorithm items; domain B consists of 4 subdomains covering 13 algorithm items; domain C consists of 4 subdomains covering 8 algorithm items; and domain D has no subdomain and covers 5 algorithm items. Our analyses focused on each of 42 algorithm items, 12 subdomains and 3 domains (A, B, and C); we did not total up domain D scores and thus did not analyze this, since this is the summary code for evidence of abnormality within the first 3 years. The assessment of domain B was further divided into two types of assessments according to verbal skills of the examined individuals; subdomains B1, B4, B2 (V), and B3 (V), covering 13 algorithm items, were used for verbal individuals, whereas only B1 and B4 were used for non-verbal individuals (including pre-speech infants).

An algorithm-based diagnosis of AD was provided if all of scores of four domains (A, B, C, and D) were equal to or exceeded the following cut-off points: 10 points for domain A; 8 points for domain BV (domain B for verbal subjects) or 7 points for domain BNV (domain B for non-verbal subjects); 3 points for domain C; and 1 point for domain D.

Reliability Study

We first calculated the weighted kappa (Kw) value for each of the 42 algorithm items; scores on the algorithm items took only one of three values (0, 1, or 2). We adopted the quadratic weighting system, that is, $w_{ij} = 1 - (i - j)^2/(k - 1)^2$ (Fleiss and Cohen 1973). This allowed Kw and the intraclass correlation coefficient (ICC) to be considered as equivalent to each other. We also calculated the ICC for each of 12 subdomains and 4 domains; the summed scores of subdomains and domains could take a number of values, and thus the ICC was preferred over the Kw. As regards judgments of the clinical level of significance, we followed the criteria provided in previous studies (Cicchetti 1994; Cicchetti and Sparrow 1981), i.e., items showing Kw $\geq$ 0.75 and subdomains/domains showing ICC $\geq$ 0.75 were regarded as excellent, 0.60 $\leq$ Kw $<$ 0.75 and 0.60 $\leq$ ICC $<$ 0.75 were considered good, and 0.40 $\leq$ Kw $<$ 0.60 and 0.40 $\leq$ ICC $<$ 0.60 were considered fair, while Kw $<$ 0.40 and ICC $<$ 0.40 exhibited poor inter-rater reliability. Considering the difference in age distribution of the three diagnostic groups of participants, analyses were first conducted on all the enrolled participants, and then a subsequent analysis was conducted separately for three age bands: below 5 years (<5:0 years); 5 years 0 months to 9 years 11 months (5:0–9:11 years); and 10 years and older.

Validity Study—Discriminant Validity

We compared the mean scores for 42 algorithm items, 12 subdomains, and 3 domains (A, B, and C) among the three diagnostic groups of participants (AD, PDDNOS, and non-ASD) using one-way ANOVA analysis with a post hoc comparison after Bonferroni's correction. We also examined whether differences in the mean scores of items, subdomains, and domains would be smaller if the analyses were limited to younger individuals (<5 years of age) or individuals exhibiting cognitive delay (IQ/DQ < 70).

Validity Study—Diagnostic Validity

To assess whether the provided diagnosis based on ADI-R-JV was diagnostically valid, we estimated the sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of ADI-R-JV. In this study, sensitivity referred to the proportion of individuals judged to have an ADI-R-JV algorithm-based diagnosis of AD among those with a consensus clinical diagnosis of AD. Specificity was the proportion of those judged not to have AD based on ADI-R-JV among those with a non-AD consensus clinical diagnosis or with no psychiatric diagnosis (i.e., subjects without a consensus clinical diagnosis of AD). PPV was the proportion of subjects with a consensus clinical diagnosis of AD among those with an algorithm-based diagnosis of AD, and NPV was the proportion of subjects with a consensus clinical diagnosis of non-AD among those with an algorithm-based diagnosis of non-AD. According to previously reported criteria (Cicchetti et al. 1995), we judged the clinical significance of sensitivity, specificity, and PPV and NPV values to be "fair" if results for these measures were equal to or exceeded 70 %, good if they were $\geq$80 %, and excellent if they were $\geq$90 %. We also examined whether results for these would be lower if the analysis were limited to that of younger individuals (<5 years of age) or individuals with an intellectual disability (IQ/DQ < 70).

Ethical Issues

The study protocol followed the ethical guidelines of the most recent Declaration of Helsinki (Edinburgh 2000) and was approved by the Institutional Ethical Review Boards at each research site. All participants, together with their caregivers, were given a complete description of the study, and the caregivers were asked to provide written informed consent to participate. As regards clinically referred individuals, they were initially contacted at one of the participating research sites, where we provided caregivers with routine feedback, which included our clinical observations and assessments. Then, by the time ADI-R-JV interview was conducted, we had formed a clinical consensus diagnosis, arrived at by experts in our research team. After ADI-R-JV interview with the caregivers had been conducted, we formulated a best-estimate diagnosis based on

both the consensus clinical diagnosis and the algorithm diagnosis. The caregivers were then provided with feedback, including a best-estimate diagnosis.

Results

Reliability Study

No single diagnostic algorithm item showed a weighted kappa (Kw) of lower than 0.6 (see Appendix Table 1 in supplementary materials). Two items showed Kw values at the level of "good" in terms of clinical significance (0.74 for item 39, "Verbal rituals", and 0.69 for item 58, "Inappropriate facial expression"), but the remaining 40 out of 42 diagnostic algorithm items showed Kw values of 0.75 or higher, indicating a level of excellent clinical significance.

All domains and subdomains showed ICC values of 0.75 or higher, indicating an excellent level (Table 3). ICC values were again calculated separately for three age bands (<5:0 years, 5:0–9:11 years, and 10–19 years). Among individuals below 5 years of age, all domains and subdomains had ICC values of ≥0.75 (excellent). For individuals between 5:0 and 9:11 years, all domains and all but one subdomain had ICC values of ≥0.75 (excellent); one exception was subdomain C3, "Stereotyped and repetitive motor mannerisms", which showed an ICC value of 0.73 (good). For those individuals 10 years of age and older, all domains and all but two subdomains showed ICC values of ≥0.75 (excellent); the exceptions were 0.69 for subdomain B2 (V), "Relative failure to initiate or sustain conversational interchange", and 0.62 for subdomain C4, "Preoccupations with part of objects or non-functional elements of material", which had ICC values over 0.6, but below 0.75 (good).

Validity Study

Discriminant Validity: Difference in Mean Scores of Items/Subdomains/Domains Across Three Diagnostic Groups

As regards the mean scores for diagnostic algorithm items (Table 4), all items but one showed a clear, significant difference across the three diagnostic groups using one-way ANOVA (AD vs. PDDNOS vs. non-ASD, $p < 0.001$

Table 3 Inter-rater reliability: intraclass correlation coefficients (ICC) of ADI-R domain and subdomain scores across three age bands (N = 51)

Domain/subdomain code	Item	ICC all subjects [N = 51]	ICC <5:0 years [N = 20]	ICC 5:0–9:11 years [N = 15]	ICC 10–19 years [N = 16]
A	Qualitative abnormalities in reciprocal social interaction	.96	.93	.97	.95
A1	Failure to use nonverbal behaviors to regulate social interaction	.92	.91	.94	.91
A2	Failure to develop peer relationships	.95	.92	.92	.90
A3	Lack of shared enjoyment	.96	.94	.98	.97
A4	Lack of socioemotional reciprocity	.91	.93	.89	.88
B	Qualitative abnormalities in communication	.97	.95	.96	.98
B1	Lack of, or delay in, spoken language and failure to compensate through gesture	.93	.94	.91	.92
B4	Lack of varied spontaneous make-believe or social imitative play	.96	.93	.97	.98
B2(V)	Relative failure to initiate or sustain conversational interchange	.92	.90	.92	.69
B3(V)	Stereotyped, repetitive, or idiosyncratic speech	.92	.96	.95	.77
C	Restricted, repetitive, stereotyped patterns of behaviour	.95	.96	.96	.87
C1	Encompassing preoccupation or circumscribed pattern of interest	.94	.97	.92	.81
C2	Apparently compulsive adherence to non-functional routines or rituals	.86	.85	.90	.81
C3	Stereotyped and repetitive motor mannerisms	.86	.85	.73	.96
C4	Preoccupations with part of objects or non-functional elements of material	.82	.89	.94	.62

Table 4 Discriminant validity: mean scores of diagnostic algorithm items, subdomains, and domains

Items	(1) AD [N = 138]	(2) PDDNOS [N = 89]	(3) Non-ASD [N = 90]	Statistics
A1. Failure to use nonverbal behaviors to regulate social interaction	3.8 (1.7)	2.6 (2.0)	0.2 (0.6)	$F(2, 314) = 138.4$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: $p < 0.001$
50. Direct gaze	1.5 (0.9)	1.1 (1.0)	0.0 (0.3)	$F(2, 227) = 61.5$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: $p = 0.002$
51. Social smiling	1.9 (1.1)	1.4 (1.2)	0.1 (0.4)	$F(2, 230) = 60.5$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: $p = 0.01$
57. Range of facial expressions used to communicate	1.2 (1.0)	0.8 (1.0)	0.0 (0.1)	$F(2, 231) = 34.1$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: $p = 0.03$
A2. Failure to develop peer relationships	5.7 (1.9)	4.4 (2.1)	0.7 (1.1)	$F(2, 314) = 226.5$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: $p < 0.001$
49. Imaginative play with peers	2.1 (0.9)	1.9 (1.0)	0.2 (0.6)	$F(2, 224) = 95.4$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: NS
62. Interest in children	1.9 (1.1)	1.4 (1.1)	0.1 (0.4)	$F(2, 229) = 63.1$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: $p = 0.01$
63. Response to approaches of other children	1.3 (0.9)	1.1 (0.8)	0.1 (0.3)	$F(2, 226) = 50.25$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: NS
64. Group play with peers	2.2 (0.8)	1.8 (0.9)	0.4 (0.7)	$F(2, 221) = 94.7$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: $p < 0.001$
65. Friendships	1.6 (1.1)	1.7 (0.9)	0.2 (0.5)	$F(2, 139) = 19.4$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: NS
A3. Lack of shared enjoyment	4.3 (1.7)	3.7 (1.8)	0.7 (1.2)	$F(2, 314) = 146.3$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: $p = 0.006$
52. Showing and directing attention	1.5 (1.2)	1.0 (1.1)	0.0 (0.3)	$F(2, 229) = 39.3$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: $p = 0.01$

Table 4 continued

Items	(1) AD [N = 138]	(2) PDDNOS [N = 89]	(3) Non-ASD [N = 90]	Statistics
53. Offering to share	2.0 (0.9)	1.7 (1.1)	0.2 (0.5)	$F(2, 227) = 75.4$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: NS
54. Seeking to share enjoyment with others	1.4 (0.7)	1.2 (0.8)	0.1 (0.3)	$F(2, 229) = 81.5$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: NS
A4. Lack of socioemotional reciprocity	6.0 (2.1)	4.3 (2.2)	0.7 (1.1)	$F(2, 314) = 226.5$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: $p < 0.001$
31. Use of other's body to communicate	1.0 (1.2)	0.8 (1.0)	0.2 (0.5)	$F(2, 273) = 226.5$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: NS
55. Offering comfort	2.1 (1.1)	1.7 (1.3)	0.0 (0.2)	$F(2, 231) = 76.5$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: NS
56. Quality of social overtures	1.7 (1.2)	1.2 (1.1)	0.1 (0.2)	$F(2, 225) = 49.9$, $p < 0.001$ 1 > 3: p < 0.001 2 > 3: $p < 0.001$ 1 > 2: $p = 0.02$
58. Inappropriate facial expression	0.9 (0.8)	0.4 (0.6)	0.0 (0.3)	$F(2, 293) = 48.5$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: $p < 0.001$
59. Appropriateness of social responses	1.7 (1.1)	1.4 (1.2)	0.2 (0.6)	$F(2, 227) = 47.1$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: $p < 0.001$
A. Quantitative abnormalities in eciprocal social interaction	19.9 (5.3)	14.8 (6.4)	2.3 (2.7)	$F(2, 314) = 330.6$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: $p < 0.001$
B1. Lack of, or delay in, spoken language and failure to compensate through gesture	4.1 (2.5)	3.0 (2.2)	0.6 (1.2)	$F(2, 314) = 79.1$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: $p < 0.001$
42. Pointing to express interest	1.2 (0.9)	0.9 (0.9)	0.1 (0.4)	$F(2, 227) = 38.4$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: NS

Table 4 continued

Items	(1) AD [N = 138]	(2) PDDNOS [N = 89]	(3) Non-ASD [N = 90]	Statistics
43. Nodding	0.9 (0.8)	0.4 (0.6)	0.0 (0.2)	$F(2, 314) = 33.8$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p = 0.01$ 1 > 2: $p < 0.001$
44. Head shaking	0.8 (0.9)	0.5 (0.8)	0.1 (0.2)	$F(2, 224) = 21.3$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p = 0.003$ 1 > 2: $p = 0.03$
45. Conventional/instrumental gesture	1.4 (1.0)	0.9 (1.0)	0.1 (0.3)	$F(2, 228) = 41.7$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: $p = 0.002$
B4. Lack of varied spontaneous make-believe or social imitative play	4.2 (1.8)	2.8 (2.0)	0.6 (1.1)	$F(2, 314) = 124.9$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: $p < 0.001$
47. Spontaneous imitation of actions	2.2 (1.1)	1.7 (1.2)	0.2 (0.6)	$F(2, 314) = 72.0$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: $p < 0.001$
48. Imaginative play	2.0 (1.1)	1.5 (1.1)	0.2 (0.6)	$F(2, 227) = 124.9$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: $p = 0.004$
61. Imitative social play	1.5 (0.9)	1.1 (1.0)	0.0 (0.1)	$F(2, 226) = 53.9$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: $p = 0.003$
B2(V). Relative failure to initiate or sustain conversational interchange	3.1 (1.3)	1.9 (1.6)	0.5 (1.1)	$F(2, 307) = 97.9$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: $p < 0.001$
34. Social verbalization/chat	1.7 (0.6)	1.4 (0.8)	0.5 (0.7)	$F(2, 314) = 67.5$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: $p = 0.01$
35. Reciprocal conversation	1.8 (0.7)	1.4 (0.8)	0.2 (0.6)	$F(2, 242) = 112.6$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: $p = 0.005$
B3(V). Stereotyped, repetitive, or idiosyncratic speech	2.9 (1.8)	2.1 (1.8)	0.9 (1.3)	$F(2, 314) = 41.2$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: $p = 0.004$
33. Stereotyped utterances and delayed echolalia	1.1 (1.1)	0.6 (0.8)	0.1 (0.4)	$F(2, 257) = 30.2$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p = 0.008$ 1 > 2: $p < 0.001$

Table 4 continued

Items	(1) AD [N = 138]	(2) PDDNOS [N = 89]	(3) Non-ASD [N = 90]	Statistics
36. Inappropriate questions or statements	1.2 (0.8)	0.6 (0.7)	0.3 (0.5)	$F(2, 258) = 45.7$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p = 0.02$ 1 > 2: $p < 0.001$
37. Pronominal reversal	0.3 (0.7)	0.1 (0.4)	0.2 (0.5)	$F(2, 221) = 2.0$, $p = 0.13$ NS
38. Neologisms/idiosyncratic language	0.4 (0.7)	0.2 (0.4)	0.2 (0.4)	$F(2, 257) = 5.9$, $p = 0.003$ 1 > 3: $p = 0.02$ 2 > 3: NS 1 > 2: $p = 0.01$
BV. Qualitative abnormalities in communications, verbal subjects	14.3 (4.1)	9.7 (4.4)	2.5 (3.2)	$F(2, 260) = 210.9$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: $p < 0.001$
BNV. Qualitative abnormalities in communications, non-verbal subjects	12.6 (4.9)	9.0 (4.4)	2.3 (2.5)	$F(2, 51) = 21.0$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.005$ 1 > 2: $p = 0.02$
C1. Encompassing preoccupation or circumscribed pattern of interest	1.9 (1.1)	0.9 (1.0)	0.3 (0.6)	$F(2, 314) = 80.6$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: $p < 0.001$
67. Unusual preoccupation	1.0 (0.9)	0.4 (0.7)	0.1 (0.3)	$F(2, 303) = 40.3$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p = 0.006$ 1 > 2: $p < 0.001$
68. Circumscribed interest	1.1 (0.8)	0.5 (0.7)	0.2 (0.5)	$F(2, 294) = 40.5$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: $p < 0.001$
C2. Apparently compulsive adherence to non-functional routines or rituals	1.4 (1.2)	0.7 (1.1)	0.2 (0.6)	$F(2, 314) = 36.3$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p = 0.01$ 1 > 2: $p < 0.001$
39. Verbal rituals	0.8 (0.9)	0.4 (0.7)	0.1 (0.3)	$F(2, 314) = 20.6$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p = 0.03$ 1 > 2: $p = 0.004$
70. Compulsions/rituals	0.9 (1.0)	0.5 (0.9)	0.2 (0.5)	$F(2, 302) = 18.1$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: NS 1 > 2: $p = 0.002$
C3. Stereotyped and repetitive motor mannerisms	0.9 (0.9)	0.5 (0.8)	0.2 (0.6)	$F(2, 314) = 19.4$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p = 0.03$ 1 > 2: $p = 0.03$

Table 4 continued

Items	(1) AD [N = 138]	(2) PDDNOS [N = 89]	(3) Non-ASD [N = 90]	Statistics
77. Hand and finger mannerisms	0.4 (0.7)	0.2 (0.5)	0.1 (0.4)	$F(2, 302) = 9.6$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: NS 1 > 2: $p = 0.004$
78. Other complex mannerisms or stereotyped body movements	0.8 (0.9)	0.4 (0.7)	0.1 (0.4)	$F(2, 303) = 21.5$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p = 0.04$ 1 > 2: $p = 0.001$
C4. Preoccupations with part of objects or non-functional elements of material	1.4 (0.7)	0.8 (0.7)	0.3 (0.6)	$F(2, 314) = 66.5$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: $p < 0.001$
69. Repetitive use of objects or interest in parts of objects	1.2 (0.9)	0.5 (0.7)	0.2 (0.4)	$F(2, 303) = 59.1$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p = 0.006$ 1 > 2: $p < 0.001$
71. Unusual sensory interests	0.7 (0.7)	0.5 (0.6)	0.2 (0.5)	$F(2, 301) = 21.7$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p = 0.006$ 1 > 2: $p = 0.006$
C. Restricted, repetitive, and stereotyped patterns of behaviors	5.5 (2.4)	2.9 (2.5)	1.1 (1.8)	$F(2, 314) = 106.6$, $p < 0.001$ 1 > 3: $p < 0.001$ 2 > 3: $p < 0.001$ 1 > 2: $p < 0.001$

NS not significant

for all items, F test); the only exception was "Pronominal reversal (item 37)" ($p = 0.13$). For the post hoc analysis, the mean scores for all items, except item 37, differed significantly between the AD and non-ASD groups. In addition, the mean scores for all items differed significantly between the PDDNOS and the non-ASD groups, with the exception of "Neologism (item 38)" ($p = 0.87$, post hoc test with Bonferroni correction); "Compulsions (item 70)" ($p = 0.15$, post hoc test with Bonferroni correction); and "Hand and finger mannerisms (item 77)" ($p = 0.22$, post hoc test with Bonferroni correction).

As regards the subdomains (A1–A4, B1–B4, C1–C4), all showed a significant difference in mean scores across the three diagnostic groups using one-way ANOVA (AD vs. PDDNOS vs. non-ASD; $p < 0.001$ for all subdomains, F test; Table 4). For the post hoc analyses, the mean of all subdomain scores revealed a significant difference between the AD and non-ASD, PDDNOS and non-ASD, and AD and PDDNOS groups.

As for domains A, B (BV/BNV), and C, the mean scores for all 3 domains were significantly different across the three diagnostic groups with one-way ANOVA (AD vs. PDDNOS vs. non-ASD; $p < 0.001$ for all domains, F test; Table 4). For the post hoc analysis, the mean scores for all domains were significantly higher in the AD than in the non-ASD group ($p < 0.001$ for domains A, BV, BNV, and C, post hoc test with Bonferroni correction), and were higher in the PDDNOS than in the non-ASD group ($p < 0.001$ for domains A, BV, and C, $p = 0.005$ for domain BNV, post hoc test with Bonferroni correction). Likewise, the mean scores of all domains were significantly higher in the AD than in the PDDNOS group ($p < 0.001$ for domains A, BV, and C, $p = 0.02$ for domain BNV, post hoc test with Bonferroni correction).

Similar comparisons of mean scores of the three domains were repeated after stratification according to three age bands (<5:0 years, 5:0–9:11 years, and 10–19 years; see Appendix Table 3 in supplementary materials). For those individuals below 5 years of age, the mean scores for all domains were significantly higher in the AD (N = 11) than in the non-ASD group (N = 45)($p < 0.001$ for domain A, $p = 0.01$ for domain

BV, $p < 0.001$ for domain BNV, $p = 0.002$ for domain C, post hoc test with Bonferroni correction), and significantly higher in the PDDNOS (N = 33) than in the non-ASD group ($p < 0.001$ for domain A and BV, $p = 0.005$ for domain BNV, $p = 0.03$ for domain C, post hoc test with Bonferroni correction). However, no significant difference was found between the AD and PDDNOS groups in any of the domains ($p = 0.19$ for domain A, $p = 0.93$ for domain BV, $p = 0.33$ for domain BNV, $p = 0.29$ for domain C, post hoc test with Bonferroni correction). As for those individuals aged 5:0–9:11 years, the mean scores of all domains (A, BV, and C; note that no group comparison was conducted in domain BNV, because there was only one nonverbal subject in the non-ASD group in this age band) were significantly higher in the AD (N = 37) than in the non-ASD group (N = 28) ($p < 0.001$ for domains A, BV, and C, post hoc test with Bonferroni correction), and were significantly higher in the PDDNOS (N = 22) than in the non-ASD group ($p < 0.001$ for domains A, BV, and C, post hoc test with Bonferroni correction). Similarly, the mean scores for all three domains were significantly higher in the AD than in the PDDNOS group ($p = 0.01$ for domains A and C, $p = 0.03$ for domain BV, post hoc test with Bonferroni correction). As for those individuals aged 10–19 years, the mean scores for all three domains (A, BV, and C; no group comparison was conducted in domain BNV, because there was only one nonverbal subject in the non-ASD group in this age band) were significantly higher in the AD (N = 90) than in the non-ASD group (N = 17) ($p < 0.001$ for domains A, BV, and C, post hoc test with Bonferroni correction). Likewise, the mean scores for all domains except domain C were higher in the PDDNOS (N = 34) and non-ASD groups ($p < 0.001$ for domains A and BV, $p = 0.07$ for domain C, post hoc test with Bonferroni correction); moreover, for all domains, mean scores were also higher in the AD than in the PDDNOS group ($p = 0.002$ for domain A, $p < 0.001$ for domain BV and C, post hoc test with Bonferroni correction).

Again, the same analyses were conducted over two groups of IQ/DQ level (<70 vs. ≥70; see Appendix Table 4 in supplementary materials). For those individuals with an IQ/DQ of <70, the mean scores for all domains (A, BV/BNV, and C) were significantly higher in the AD (N = 18) than in the non-ASD (N = 8) group ($p < 0.001$ for domains A and C, $p = 0.007$ for domain BV and $p = 0.05$ for domain BNV, post hoc test with Bonferroni correction). The mean scores for domains A and BV were significantly higher in the PDDNOS (N = 9) than in the non-ASD group (N = 8), but no significant difference was found for domains BNV, and C ($p < 0.001$ for domain A, $p = 0.05$ for domain BV, $p = 0.13$ for domain BNV, $p = 0.99$ for domain C, post hoc test with Bonferroni correction). A significant difference in mean scores between the AD and PDDNOS groups was found only in domain C ($p = 0.99$ for domain A, $p = 0.08$ for domain BV, $p = 0.99$ for domain BNV, $p < 0.001$ for domain C, post hoc test with Bonferroni correction). In turn, for those individuals with an IQ/DQ of ≥70, mean scores for all domains were significantly higher in the AD (N = 120) than in the non-ASD group (N = 82) ($p < 0.001$ for domains A, BV, BNV, and C, post hoc test with Bonferroni correction), higher in the PDDNOS (N = 80) than in the non-ASD group ($p < 0.001$ for domains A, BV, C, $p = 0.002$ for domain BNV, post hoc test with Bonferroni correction), and higher in the AD than in the PDDNOS group ($p < 0.001$ for domains A, BV, C, $p = 0.01$ for domain BNV, post hoc test with Bonferroni correction).

Diagnostic Validity: Agreement with Consensus Clinical Diagnosis of AD

In our analysis of the overall diagnostic validity of the Japanese version of ADI-R, we found that across all individuals, the sensitivity, specificity, PPV, and NPV of the test were very high (92, 89, 87, and 93 %, respectively; Table 5). Similar results were also obtained for age groups 5:0–9:11 years and

Table 5 Diagnostic validity: agreement with consensus clinical diagnosis among those with algorithm diagnosis of AD

	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
Consensus clinical diagnosis: *Autistic disorder* [N = 138]				
Algorithm diagnosis of AD: Domain A ≥ 10 AND (Domain BV ≥ 8 for verbal OR BNV ≥ 7 for non-verbal subjects) AND Domain C ≥ 3 AND Domain D ≥ 1 (Rutter et al. 2003)				
All individuals [N = 317]	92	89	87	93
Age: below 4:0 [N = 73]	53	92	55	92
Age: below 5:0 [N = 89]	55	92	50	93
Age: 5:0–9:11 [N = 87]	92	84	81	93
Age: 10:0 and older [N = 141]	97	90	95	94
IQ/DQ: below 70 [N = 35]	94	100	100	94
IQ/DQ: 70 and over [N = 282]	92	88	85	93

Sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV)

older, and for IQ/DQ groupings below 70 and at 70 and above. Consistent with our initial hypotheses, the sensitivity and PPV for ages below 4 and below 5 years were both poor, i.e., between 50 and 55 %, respectively.

Overall test sensitivity, or the proportion of individuals with AD (N = 138) who were correctly categorized as having AD using ADI-R-JV, was as high as 92 %, indicating excellent clinical significance, which was also shown for the evaluation of individuals aged 5:0–9:11 and age 10 years and older, and for those individuals at either cognitive level assessed, i.e., with a score of <70 or ≥70. However, for individuals aged below 5 years, a sensitivity of 55 % was found, indicating a poor level of clinical significance.

On the other hand, among individuals without a consensus clinical diagnosis of AD (N = 179), 159 were also judged not to have AD based on ADI-R-JV algorithm diagnosis, i.e., the specificity of ADI-R-JV for correctly excluding AD was 89 % (159/179), indicating excellent clinical significance. This clinically excellent specificity was replicated for individuals in each of the three age bands, and for both IQ/DQ bands examined.

Discussion

In the present study, we reported the inter-rater reliability, discriminant validity, and diagnostic validity of the Japanese Version of ADI-R (ADI-R-JV).

Reliability of ADI-R-JV

In agreement with our hypotheses, the Kw values for all algorithm items of ADI-R-JV exceeded a value of 0.6, which was also consistent with the findings of previous studies (Hill et al. 2001; Lord et al. 1994). Furthermore, among the 42 algorithm items, all but two (items 39 and 58) showed Kw values in excess of 0.75, indicating excellent inter-rater reliability; the two exceptions showed Kw values of 0.60–0.75, indicating good inter-rater reliability.

We also investigated whether the measures for inter-rater reliability would decrease when the analysis was limited to individuals in specific age bands (Table 3). Again, the ICCs for all domains and subdomains exceeded 0.75 (excellent) among individuals aged less than 5 years, and the ICCs for all but 1 (C3) subdomain exceeded 0.75 (excellent) among individuals aged 5:0–9:11 years. Of note, ICCs can be seen as reflecting a good to excellent level of clinical significance, regardless of the age of the examinee. It is worth mentioning in this context that the ICCs became smaller in subdomains B2(V), B3(V), C1, C2, and C4 if the examinees were 10 years old or older. This tendency, i.e., smaller ICC values of the age band of 10 years and older, should first be discussed in light of the definition of inter-rater reliability, which can be easily compromised when the degree of experience and training of pairs conducting the interviews differs. When such a difference in experience occurred in the present study, compromised ICCs should have been observed irrespective of a subject's age, since the two raters were selected on a random basis from each site. Furthermore, the raters who administered ADI-R-JV were fully and equally experienced after the official training sessions. Therefore, the compromised ICCs for those subjects 10 years old and older did not seem to reflect a bias stemming from assessment skills. There is agreement between our findings and previous results showing lower scores for items under domain C than for items under domains A and B (Hill et al. 2001; Lord et al. 1994). Specifically, the inter-rater reliability of items under domain C would be particularly likely to be compromised when the examinees were older (i.e., 10 years and older), probably due to the uncertain recall of remote episodes. However, since we only obtained limited findings regarding inter-rater reliability upon assessment of adolescent subjects, elaboration on this topic remains difficult.

On the other hand, ICCs were not lower when the analysis was limited to the examination of individuals with an intellectual disability (IQ/DQ < 70), or when only males or females were included in the analysis (Table not shown). Rather, under no circumstances did we observe a Kw or ICC below 0.6 (Table 3). These findings strongly indicate the satisfactory inter-rater reliability of ADI-R-JV, i.e., the translated version appears to be as reliable as the original ADI-R in English.

Validity of ADI-R-JV

Discriminant Validity

Mean scores for three domains (A, B[BV/BNV], and C) were significantly higher in the AD group than in the PDDNOS and the non-ASD groups, indicating that the discriminant validity of ADI-R-JV was stable. Thus, our results appear to be consistent with the findings of previous pivotal studies investigating younger individuals with AD (Lord et al. 1993; Saemundsen et al. 2003), even in those with concomitant developmental delay (Gray et al. 2008).

Originally, ADI-R was designed to detect AD, not ASD (Lord et al. 1994). Therefore, in the current analysis, we expected not only that the mean scores for all domains would be higher in the AD group than in the non-ASD group, but also that they would be higher in the AD than in the PDDNOS group. These two hypotheses held true when the analysis included all study participants (N = 317).

However, the latter hypothesis (mean scores for AD > mean scores for PDDNOS) did not hold true when the analysis was limited to individuals less than 5 years of age (Appendix Table 3 in supplementary materials). Presumably, one of the main reasons for the compromised discriminability (i.e., no difference between AD and PDDNOS reflected in ADI-R-JV scores for younger individuals) was that it is difficult to differentiate AD from PDDNOS in individuals younger than 5 years of age (Turner and Stone 2007). On the other hand, the present finding may also have been due to biases; for instance, the diagnostic algorithms were prepared separately for those aged 4 years and older (based on current and past behavior) and for those younger than 4 years of age (based on current behavior). Thus, it would be possible that the discriminant validity would differ for individuals younger than 4 years old and for individuals between 4 and 5 years old. We thus analyzed a restricted sample of individuals below 4 years of age (N = 73), and found that the mean scores for domain A were 14.5 for AD, 11.4 for PDDNOS, and 3.1 for non-ASD. These results indicated that the mean was slightly higher in the AD group than in the PDDNOS group ($p = 0.051$, after Bonferroni correction), whereas the mean scores for domain BV/BNV and domain C did not reveal such differences between the AD and PDDNOS groups, suggesting that the choice of algorithm according to age may have at least partly affected the results for younger individuals.

As regards to the above results stratified by age, attention should be paid to our sample selection; among individuals below 5 years of age, 12 % had AD and 33 % had PDDNOS, whereas 64 % had AD and 24 % had PDDNOS among individuals who were 10 years old or older. These figures are consistent with differences in mean age across the three diagnostic groups shown in Table 2, and that a sample bias influenced the results. If we were to have recruited younger children with AD in the analysis, a higher level of discrimination among subgroups would likely have been observed.

Discriminant validity was also compromised for individuals with an intellectual disability (IQ/DQ < 70, see Appendix Table 4 in supplementary materials). Again, we expected that the mean scores for all domains were higher in the AD group than in both the non-ASD and PDDNOS groups. The first hypothesis (mean scores for AD > mean scores for non-ASD) held true for all domains, regardless of IQ level. However, the second hypothesis (mean scores for AD > mean scores for PDDNOS) held true only for domain C among individuals with an IQ/DQ of < 70; instead, the relationship of mean scores for PDDNOS > mean scores for non-ASD was not observed for domain C among individuals with an IQ/DQ of < 70. These results suggest that the relevance of domain C in arriving at a diagnosis of AD may differ from the relevance of domains A and B, particularly for individuals with a developmental delay. This issue has already been addressed in the literature; some authors have argued that the exclusion of domain C may improve discriminability between toddlers with and without ASD (Ventola et al. 2006). Furthermore, Lord and Jones (2012) reviewed that compared to symptoms under the social interaction and communication domains, symptoms under the repetitive behavior domain (domain C) are more heterogeneous across individuals and context-dependent, and thus caregivers may not consistently notify clinicians about domain-C symptoms. Our findings appear to be in line with the results of these previous studies. Specifically, individuals with a consensus diagnosis of AD with concomitant cognitive delay would be diagnosed as having Social Communication Disorder according to the proposed version in the DSM-5 (http://www.dsm5.org/ProposedRevision/Pages/NeurodevelopmentalDisorders.aspx), using ADI-R-JV. This issue still needs to be addressed in future studies.

Thus far, the overall discriminant validity of ADI-R-JV has been shown to be sufficient, although it appeared compromised for the assessment of younger individuals and individuals with concomitant cognitive delay. Potential biases and the limited statistical power of the present study should be noted, as these factors might have resulted in the finding of compromised discriminability among younger individuals.

As shown in Table 4, "Pronominal reversal (item 37)" showed no statistical difference among the three diagnostic groups. This finding was of interest in terms of language use, because in Japanese conversations, personal pronouns are not as frequently used as they are in English. In addition, even when personal pronouns are not used, there are no verbal conjugations in Japanese that correspond to those in Latin-derived languages. We are certain that this specific feature of the Japanese language allowed the mean scores on item 37 to remain fairly close to zero. Nevertheless, this concern did not in any way affect discriminability among domain scores, nor was diagnostic validity affected.

Diagnostic Validity

The sensitivity of ADI-R-JV with respect to correctly diagnosing autistic disorder was 92 %, indicating that the overall sensitivity of the instrument is excellent. Moreover, the algorithm's overall specificity, which was shown to be 89 %, was determined to be good. Likewise, the overall PPV and NPV were 89 and 93 %, respectively, indicating good to excellent clinical significance, consistent with our expectations. These figures were similar or even better than those obtained in a recent study using a translated version of ADI-R administered to individuals with a mean age of 10 years (Lampi et al. 2010). However, in the current study, the corresponding sensitivity decreased to 55 % (indicating

poor sensitivity; Table 5) when the analysis was limited to subjects younger than 5 years of age, suggesting that diagnostic validity was compromised in younger individuals. This finding was also consistent with our hypothesis. The compromised sensitivity for younger individuals may be rather straightforward; prior studies have been consistent with this finding, and our own results indicated compromised discriminability between AD and PDDNOS individuals below 5 years of age. However, as such compromised discriminability was not firmly upheld due to potential biases and the limited statistical power of our study sample, analysis of a larger number of individuals may have provided a higher level of sensitivity. Indeed, a recent large-scale study indicated a sensitivity for correctly diagnosing AD as high as 82.7 %, even when participants were under the age of 36 months (Risi et al. 2006). Nevertheless, it remains possible that the low level of sensitivity for those aged less than 5 years in the present study was not simply due to sample selection or the algorithm applied, but also a reflection of the difficulty of differentiating AD from PDDNOS in individuals at such young age, as was suggested by recent literature (Turner and Stone 2007).

In light of the proposed diagnosis of ASD in the forthcoming Diagnostic and Statistical Manual of Mental Disorders (version 5), research interests have increasingly focused on differentiating ASD from non-ASD individuals using ADI-R; however, there is no established cutoff for ASD in ADI-R. Attempts have been made to apply the original algorithm to ASD individuals; unfortunately, sensitivity for correctly diagnosing ASD was shown to be insufficient (Kim and Lord 2012; Risi et al. 2006). A related attempt to differentiate ASD from non-ASD individuals using ADI-R was the use of other assessment scales such as the Vineland Adaptive Behavior Scale (Sparrow et al. 1984) to improve sensitivity (Tomanik et al. 2007). Another attempt at differentiation was to relax the original, stringent algorithm for AD. For instance, in one genetic study (International Molecular Genetic Study of Autism Consortium 2001), the diagnosis of ASD was made according to ADI-R, whereby exceeding the cutoffs of three domains (A, B, C) was required for ASD diagnosis, with the exception that a score on any one of the three domains could fall one point below the threshold. We recalculated sensitivity using this relaxed criterion in the current study, resulting in an overall sensitivity of 64 %. When the same analysis was repeated for three age bands, sensitivity was 27 % for subjects aged < 5 years old, 71 % for subjects aged 5:0–9:11 years old, and 74 % for those 10 years old and older (Table not shown). At present, ADI-R-JV appears to have limited diagnostic validity with respect to detecting ASD.

Nevertheless, studies have emphasized that the use of ADOS together with ADI-R is a sensible approach, in that the combination of the two reflects consensus clinical judgments of AD as well as of ASD better than any other single instrument used alone (Le Couteur et al. 2008), even in individuals as young as 3 years old and younger (Risi et al. 2006). In this regard, evaluations of the sensitivity of both the Japanese version of ADOS and ADI-R-JV for correctly diagnosing ASD should be conducted.

It should also be noted that the sensitivity of ADI-R-JV with respect to correctly diagnosing AD among individuals with concomitant cognitive delay (IQ/DQ < 70) was 94 %, i.e., not lower than the corresponding result for individuals with an IQ of >70 (92 %); this findings was inconsistent with our expectations, as well as with a prior study (de Bildt et al. 2004). Furthermore, other studies have shown that specificity was more prone than sensitivity to be compromised when the examinee exhibited cognitive delay, and thus individuals with cognitive delay are more likely to be overdiagnosed (Lord et al. 1994; Risi et al. 2006). As regards the discrepancy with our hypothesis, the sample bias of the present study should be taken into account, because the mean IQ/DQ of individuals with AD and PDDNOS in this study was fairly high, even higher than reported in previous studies. In addition, the small number of enrolled participants with an IQ/DQ of <70 could have limited the statistical power of the study to detect any compromising effects of cognitive delay on diagnostic validity.

Limitations and Strengths

Treatment or interventions that may have affected the children enrolled in this study should also be taken into account, particularly in the assessment of diagnostic subgroups. It was a limitation of this study that we did not collect relevant data on this topic. On the other hand, ADI-R is a measure based principally on the observation of past behavior during early stages of development, and usually is employed prior to such interventions, and is not based on a patient's current status. This means that the scores we obtained were less likely to reflect intervention effects compared to the scores of instruments that assess current behaviors, such as ADOS. In addition, we observed good to excellent inter-rater reliability, discriminant validity, and diagnostic validity of ADI-R-JV even without considering treatment effects that would have been observed among clinically referred individuals. Considering that statistical tests are generally biased toward null hypotheses (no difference), an adjustment allowing for treatment effects, when examined, would increase the validity of the ADI-R-JV.

In the present study, clinically referred and control individuals were enrolled according to different protocols. If caregiver motivation to participate in this study differed

for the two groups of individuals examined, the difference may have been a substantial source of sample bias. The most likely scenario related to this issue would be that a caregiver of a control individual was highly motivated to participate in the study when there was a concern that the child may have had an undiagnosed psychiatric disorder such as ASD. Indeed, such motivation might have been reflected in high proportions of non-ASD psychiatric disorders; 2 out of 16 control individuals in the reliability study (Table 1) and 4 out of 82 control individuals had such a diagnosis. Parental education and socioeconomic status, when available, may have provided some insight into the extent of this problem, but unfortunately we did not collect such data, which might otherwise have helped to refute this scenario. However, if such a motivation to participate in the study had indeed been the case, it is likely that a number of individuals with ASD would have been detected among control individuals, yet there was not a single case of undiagnosed ASD (i.e., later detected as such) among individuals initially enrolled as controls (Table 1 and Appendix Table 2 in supplementary materials). To minimize this ambiguity, confirmatory studies will be necessary.

Consensus clinical diagnoses were obtained through clinical assessments and case reviews of all of the available information, albeit outside the context of the administration of ADI-R-JV. This approach might have led to a lack of information for optimizing the diagnosis, but it ensured the independence of the administration of the ADI-R-JV. Moreover, ADI-R-JV was administered in a blinded fashion without any reference to the clinical consensus diagnosis, which could also be considered as a strength of the present study.

When we finalized our consensus clinical diagnosis, it might have been helpful to facilitate diagnosis derived from ADOS. It may also have been helpful to adopt this protocol as an external criterion for estimating the validity of ADI-R-JV. Indeed, the Japanese translation of ADOS has been available to those who established the research reliability of ADOS (i.e., since 2010). Our research team consists of very experienced clinicians and clinical researchers, and among the 8 team members involved in establishing a consensus clinical diagnosis, 4 had already established, and 2 were planning to establish, the research reliability of the ADI-R; 3 had already established, and 3 were planning to establish, the research reliability of ADOS; and each member had participated in at least one research training session on either ADI-R or ADOS. Thus, all the team members involved in establishing a consensus clinical diagnosis were fully knowledgeable about the current diagnosis of ASD in a research setting.

Conclusions

ADI-R-JV is a reliable tool, and has sufficient ability to discriminate between individuals with AD and other diagnoses, as well as between individuals with AD and those with no psychiatric diagnosis. The sensitivity for correctly diagnosing AD was generally high (92 %), but appeared to be compromised (55 %) when the tool was used to assess children younger than 5 years of age. The specificity of ADI-R-JV was consistently high, regardless of the age and cognitive level of the examinee.

Acknowledgments Declaration of interest: None of the authors have any conflict of interest to declare. The authors thank Dr. Masahiro Oshima, M.D., for recruiting children with ASD. Funding of this study was provided by a Research Grant from the Ministry of Health, Labour and Welfare of Japan (H20-KOKORO-004: K.J. Tsuchiya, Y. Kamio), by a grant for the National Center for Child Health and Development (21S-3: K. J. Tsuchiya), by the Pfizer Health Research Fund (K. Matsumoto), and by Grants-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan (A) (No. 20591396: N. Takei). These funding sources had no role in the study design; in the collection, analysis and interpretation of the data; in the writing of the report; or in the decision to submit the paper for publication.

References

American Psychiatric Association. (2000). *Task force on DSM-IV. Diagnostic and statistical manual of mental disorders: DSM-IV-TR*. Washington, DC: American Psychiatric Association.

Baron-Cohen, S., Scott, F. J., Allison, C., Williams, J., Bolton, P., Matthews, F. E., et al. (2009). Prevalence of autism-spectrum conditions: UK school-based population study. *British Journal of Psychiatry, 194*, 500–509.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284–290.

Cicchetti, D. V., Lord, C., Koenig, K., Klin, A., & Volkmar, F. R. (2008). Reliability of the ADI-R: Multiple examiners evaluate a single case. *Journal of Autism and Developmental Disorder, 38*, 764–770.

Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *American Journal of Mental Deficiency, 86*, 127–137.

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

Cox, A., Klein, K., Charman, T., Baird, G., Baron-Cohen, S., Swettenham, J., et al. (1999). Autism spectrum disorders at 20 and 42 months of age: stability of clinical and ADI-R diagnosis. *Journal of Child Psychology and Psychiatry, 40*, 719–732.

de Bildt, A., Sytema, S., Ketelaars, C., Kraijer, D., Mulder, E., Volkmar, F., et al. (2004). Interrelationship between autism diagnostic observation schedule-generic (ADOS-G), autism diagnostic interview-revised (ADI-R), and the diagnostic and statistical manual of mental disorders (DSM-IV-TR)

classification in children and adolescents with mental retardation. *Journal of Autism and Developmental Disorder, 34*, 129–137.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurements, 33*, 613–619.

Fombonne, E. (2009). Epidemiology of pervasive developmental disorders. *Pediatric Research, 65*, 591–598.

Gilchrist, A., Green, J., Cox, A., Burton, D., Rutter, M., & Le Couteur, A. (2001). Development and current functioning in adolescents with Asperger syndrome: a comparative study. *Journal of Child Psychology and Psychiatry, 42*, 227–240.

Gray, K. M., Tonge, B. J., & Sweeney, D. J. (2008). Using the autism diagnostic interview-revised and the autism diagnostic observation schedule with young children with developmental delay: Evaluating diagnostic validity. *Journal of Autism and Developmental Disorder, 38*, 657–667.

Hill, A., Bolte, S., Petrova, G., Beltcheva, D., Tacheva, S., & Poustka, F. (2001). Stability and interpersonal agreement of the interview-based diagnosis of autism. *Psychopathology, 34*, 187–191.

International Molecular Genetic Study of Autism Consortium. (2001). A genomewide screen for autism: Strong evidence for linkage to chromosomes 2q, 7q, and 16p. *American Journal of Human Genetics, 69*, 570–581.

Kawamura, Y., Takahashi, O., & Ishii, T. (2008). Reevaluating the incidence of pervasive developmental disorders: impact of elevated rates of detection through implementation of an integrated system of screening in Toyota, Japan. *Psychiatry and Clinical Neurosciences, 62*, 152–159.

Kim, Y. S., Leventhal, B. L., Koh, Y. J., Fombonne, E., Laska, E., Lim, E. C., et al. (2011). Prevalence of autism spectrum sisorders in a total population sample. *American Journal of Psychiatry, 168*, 904–912.

Kim, S. H., & Lord, C. (2012). New autism diagnostic interview-revised algorithms for toddlers and young preschoolers from 12 to 47 months of age. *Journal of Autism and Developmental Disorder, 42*, 82–93.

Kočovská, E., Biskupstø, R., Carina Gillberg, I., Ellefsen, A., Kampmann, H., Stora, T., et al. (2012). The rising prevalence of autism: A prospective longitudinal study in the Faroe Islands. *Journal of Autism and Developmental Disorder,*. doi: 10.1007/s10803-012-1444-9.

Koyama, T., Osada, H., Tsujii, H., & Kurita, H. (2009). Utility of the Kyoto scale of psychological development in cognitive assessment of children with pervasive developmental disorders. *Psychiatry and clinical neurosciences, 63*, 241–243.

Lampi, K. M., Sourander, A., Gissler, M., Niemela, S., Rehnstrom, K., Pulkkinen, E., et al. (2010). Validity of finnish registry-based diagnoses of autism with the ADI-R. *Acta Paediatrica, 99*, 1425–1428.

Le Couteur, A., Haden, G., Hammal, D., & McConachie, H. (2008). Diagnosing autism spectrum disorders in pre-school children using two standardised assessment instruments: The ADI-R and the ADOS. *Journal of Autism and Developmental Disorder, 38*, 362–372.

Lord, C., & Jones, R. M. (2012). Annual research review: Re-thinking the classification of autism spectrum disorders. *Journal of Child Psychology and Psychiatry, 53*, 490–509.

Lord, C., Risi, S., DiLavore, P. S., Shulman, C., Thurm, A., & Pickles, A. (2006). Autism from 2 to 9 years of age. *Archives of General Psychiatry, 63*, 694–701.

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Jr., Leventhal, B. L., DiLavore, P. C., et al. (2000). The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorder, 30*, 205–223.

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 Disorder, 24*, 659–685.

Lord, C., Storoschuk, S., Rutter, M., & Pickles, A. (1993). Using the ADI-R to diagnose autism in preschool children. *Infant Mental Health Journal, 14*, 234–252.

McGovern, C. W., & Sigman, M. (2005). Continuity and change from early childhood to adolescence in autism. *Journal of Child Psychology and Psychiatry, 46*, 401–408.

Mildenberger, K., Sitter, S., Noterdaeme, M., & Amorosa, H. (2001). The use of the ADI-R as a diagnostic tool in the differential diagnosis of children with infantile autism and children with a receptive language disorder. *European Child and Adolescent Psychiatry, 10*, 248–255.

Parner, E. T., Schendel, D. E., & Thorsen, P. (2008). Autism prevalence trends over time in Denmark: Changes in prevalence and age at diagnosis. *Archives of Pediatrics and Adolescent Medicine, 162*, 1150–1156.

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

Rutter, M., Le Couteur, A., & Lord, C. (2003). *ADI-R: Autism diagnostic interview revised. WPS edition manual.* Los Angeles, CA: Western Psychological Services.

Saemundsen, E., Magnússon, P., Smári, J., & Sigurdardóttir, S. (2003). Autism diagnostic interview-revised and the childhood autism rating scale: Convergence and discrepancy in diagnosing autism. *Journal of Autism and Developmental Disorder, 33*, 319–328.

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

Tanaka Institute of Education. (1987). *Tanaka-Binet intelligence scale.* Tokyo: Taken Publisher.

Tomanik, S. S., Pearson, D. A., Loveland, K. A., Lane, D. M., & Bryant Shaw, J. (2007). Improving the reliability of autism diagnoses: Examining the utility of adaptive behavior. *Journal of Autism and Developmental Disorder, 37*, 921–928.

Turner, L. M., & Stone, W. L. (2007). Variability in outcome for children with an ASD diagnosis at age 2. *Journal of Child Psychology and Psychiatry, 48*, 793–802.

Ventola, P. E., Kleinman, J., Pandey, J., Barton, M., Allen, S., Green, J., et al. (2006). Agreement among four diagnostic instruments for autism spectrum disorders in toddlers. *Journal of Autism and Developmental Disorder, 36*, 839–847.

Waterhouse, L. (2008). Autism overflows: Increasing prevalence and proliferating theories. *Neuropsychology Review, 18*, 273–286.

Wechsler, D., Golombok, S., & Rust, J. (1992). *Wechsler intelligence scale for children—Third edition (WISC-IIIUK).* London, UK: The Psychological Corporation.

Williams, J. G., Higgins, J. P., & Brayne, C. E. (2006). Systematic review of prevalence studies of autism spectrum disorders. *Archives of Disease in Childhood, 91*, 8–15.

自立に向けて
■
学校でできる支援

通常学級で特別支援を進めるために

中京大学教授　辻井正次（つじいまさつぐ）

はじめに

今回、通常学級での特別支援というタイトルの中でこの小論を進めていきます。すでに読者の皆さんは理解しておられるように、二〇〇五年から施行されている発達障害者支援法以降、発達障害のある子どもに対しての適切な教育的支援は法的に義務付けられており、障害特性に対応した「合理的な配慮」をしない場合、今後は法的責任を問われる場合も出てくることが予想されています。ですので、むしろ特別支援教育は通常学級を主戦場として行っていくもの、というふうに、教育現場という支援の舞台の構成は組み替わっていま す。わが国は法治国家ですので、法的に義務付けられたという意味は、教員の個人的な判断でやらなくていいというようなものではなく、すべての教員が当たり前に取り組むものだということはまず理解しておく必要があります。

特別支援は、別に特別な支援ではない

そもそも最初に考えておかなければならないのは、「発達障害」であろうがなかろうが、つまり、発達障害とされている自閉症スペクトラム障害や注意欠陥多動性障害や学習障害の診断があろうがなかろうが、発達障害の「特性」をもつ場合には、教員は特別支援を