Eur Child Adolesc Psychiatry (2016) 25:769–780
DOI 10.1007/s00787-015-0793-2

ORIGINAL CONTRIBUTION

# The objectivity of the Autism Diagnostic Observation Schedule (ADOS) in naturalistic clinical settings

**Eric Zander[1,2,3] · Charlotte Willfors[1,2] · Steve Berggren[1,2] · Nora Choque-Olsson[1,2,4] · Christina Coco[1,2,5] · Anna Elmund[6,7] · Åsa Hedfors Moretti[8] · Anette Holm[9] · Ida Jifält[10,11] · Renata Kosieradzki[12,13] · Jenny Linder[14] · Viviann Nordin[5] · Karin Olafsdottir[15] · Lina Poltrago[6] · Sven Bölte[1,2]**

Received: 30 June 2015 / Accepted: 29 October 2015 / Published online: 19 November 2015


**Abstract** The Autism Diagnostic Observation Schedule (ADOS) is a first-choice diagnostic tool in autism spectrum disorder (ASD). Excellent interpersonal objectivity (interrater reliability) has been demonstrated for the ADOS under optimal conditions, i.e., within groups of highly trained "research reliable" examiners in research setting. We investigated the spontaneous interrater reliability among clinically trained ADOS users across multiple sites in clinical routine. Forty videotaped administrations of the ADOS modules 1–4 were rated by five different raters each from a pool of in total 15 raters affiliated to 13 different clinical sites. $G(q,k)$ coefficients (analogous to intraclass correlations), kappas ($\kappa$) and percent agreement (PA) were calculated. The median interrater reliability for items across the four modules was $G(q,k) = .74–.83$, with the single ADOS items ranging from .23 to .94. $G(q,k)$ for total scores was .85–.92. For diagnostic classification (ASD/non-spectrum), PA was 64–82 % and Fleiss' $\kappa$ .19–.55. Objectivity was lower for pervasive developmental disorder not otherwise specified and non-spectrum diagnoses as compared to autism. Interrater reliabilities of the ADOS items and domain totals among clinical users across multiple sites were in the same range as previously reported for research reliable users, while the one for diagnostic classification was lower. Differences in sample characteristics, rater skills and statistics compared with previous studies are discussed. Findings endorse the objectivity of the

**Electronic supplementary material** The online version of this article (doi:10.1007/s00787-015-0793-2) contains supplementary material, which is available to authorized users.

✉ Eric Zander
eric.zander@ki.se

1 Pediatric Neuropsychiatry Unit, Department of Women's and Children's Health, Center of Neurodevelopmental Disorders (KIND), Karolinska Institutet, Gävlegatan 22B, 113 30 Stockholm, Sweden

2 Child and Adolescent Psychiatry, Center for Psychiatry Research, Stockholm County Council, Stockholm, Sweden

3 Neurodevelopmental Psychiatry Unit South East, Child and Adolescent Psychiatry, Stockholm County Council, Stockholm, Sweden

4 BUP Södertälje, Child and Adolescent Psychiatry, Stockholm County Council, Södertälje, Sweden

5 Neuropediatric Unit, Sachs' Children and Youth Hospital, Stockholm County Council, Stockholm, Sweden

6 PRIMA Child and Adolescent Psychiatry, Stockholm, Sweden

7 Citypsykologhus, Stockholm, Sweden

8 BUP Sollentuna, Child and Adolescent Psychiatry, Stockholm County Council, Sollentuna, Sweden

9 Astrid Lindgrens Children's Hospital, Karolinska University Hospital Solna, Stockholm County Council, Solna, Sweden

10 Astrid Lindgrens Children's Hospital, Karolinska University Hospital Huddinge, Stockholm County Council, Huddinge, Sweden

11 Pupil Health Unit, Tiohundra, Norrtälje, Sweden

12 BUP Malmö, Child and Adolescent Psychiatry, Region Skåne, Malmö, Sweden

13 Pupil Health Unit, Resource Team for Learning Disabled, City of Malmö, Sweden

14 Södra Älvsborg Hospital (SÄS), Child and Adolescent Psychiatry, Region Västra Götaland, Borås, Sweden

15 BUP Lund, Child and Adolescent Psychiatry, Region Skåne, Lund, Sweden

ADOS in naturalistic clinical settings, but also pinpoint its limitations and the need and value of adequate and continuous rater training.

**Keywords** Autism spectrum disorder · Interrater reliability · Diagnostic instrument

## Introduction

"Best clinical judgment of experienced clinicians" is still considered the gold standard of diagnosing autism spectrum disorder (ASD) [1–4]. However, even for experienced clinical experts the use of standardized diagnostic instruments within the framework of a multidisciplinary comprehensive diagnostic assessment is recommended as a potential means to improve individual diagnostic decision making over time and across clinicians [5, p 55]. The Autism Diagnostic Observation Schedule (ADOS/-2) [6, 7] (in the following the acronym ADOS will be used for both versions) and the compatible Autism Diagnostic Interview-Revised (ADI-R) [8] are among the most widely used diagnostic instruments in research as well as in everyday clinical practice around the world [9–11]. In child and adolescent psychiatry in the USA, Europe and other countries, the ADOS has become a first-choice diagnostic instrument, often designated as "gold standard." Its popularity is probably owing to its clinicalness. Reminding of common clinical observation setting, it samples and quantifies an individual's behaviors during a naturalistic semi-structured direct observation and provides a clear-cut diagnostic classification as well as severity scores to support categorical and dimensional decision making. In addition, the ADOS is one of the best psychometrically evaluated and conceptually updated tools in ASD assessment, covering a large part of ASD presentations. The availability of standardized diagnostic instruments such as the ADOS opens up for the expectation that even less experienced clinicians should be capable of accurately diagnosing ASD, when applying these tools. Indeed, in research the ADOS diagnostic classification has sometimes been used as a proxy or necessary criterion for diagnosis [12–14].

Many studies have reported excellent diagnostic validity of the ADOS, mostly from research settings [15–27]. Psychometrically, a necessary prerequisite for diagnostic validity is objectivity. Objectivity is the degree to which a scale is independent of outside influences, such as a specific administrator conducting the assessment. Studying objectivity might be particularly relevant to the ADOS, as it poses high demands on the examiner to fulfill the needs of standardization in face of a complex and unpredictable course of administration. Objectivity in terms of interpersonal objectivity is psychometrically often operationalized as interrater reliability and/or interrater agreement. The objectivity of ADOS ratings is often discussed among clinicians, for example during clinical training, where the coding procedure sometimes is perceived as arbitrary by beginners and experienced users. The authors of the ADOS are well aware of the significance of objectivity, why they have established educational standards for the use of the instrument in research settings. In this regard, becoming so called research reliable, demands specific, advanced training and prolonged hands-on supervision to achieve at least 80 % exact agreement in the coding of all ADOS module items on several occasions with a (research reliable) certified ADOS trainer. In the course of the development of the ADOS [6, 7, 28, 29], it was shown that within a group of research reliable ADOS examiners interrater reliability is excellent.

For instance, Lord et al. [28] demonstrated that it was possible for groups of well-prepared raters to reach a substantial level of interrater reliability for items of the pre-published versions of the ADOS ($\kappa_w$ [weighted kappa] between .58 and .92) and the PL-ADOS ($\kappa_w = .60–1.00$) [29], see Online resources Table 1 for details. The first published version of the ADOS contained the most elaborated interrater reliability study of the ADOS to date [6]. The interrater reliability of 12 raters' assessments of 98 individuals ($n = 20–29$ per module) on item level was analyzed using $\kappa_w$ and percent (exact) agreement (PA). For module 1, all but one item had a $\kappa_w > .60$ ($\kappa_w = .55–1.00$, median = .78), most items of module 2 had a $\kappa_w > .50$ ($\kappa_w = .38–.93$, median = .65), and many items of modules 3 and 4 had a $\kappa_w > .60$ (module 3: $\kappa_w = .46–1.00$, median = 61; module 4: $\kappa_w = .41–.93$, median = 60). Intraclass correlation (ICC) was used to assess the interrater reliability of the domain totals for this sample. Pooled for all modules ($n = 97$), the ICC was .93 for social interaction, .84 for communication, .92 for social interaction and communication combined and .82 for stereotyped behaviors and restricted interests. PA was used for classification: module 1: 93 %, module 2: 87 %, module 3: 81 % and module 4: 84 % exact agreement when all participants (both with autistic disorder and pervasive developmental disorder not otherwise specified [PDD-NOS]) were included, but PA was higher if only individuals with autism and NS were included (90–100 %). Additional interrater reliability data (another subsample of the dataset of 1999/2000 and the same 12 raters) for domain totals and classification of the revised algorithms were published in the ADOS-2 manual [7]. The ICC of the social affect (SA) domain was .97 for module 1 ($n = 63$), .98 for module 2 ($n = 50$) and .92 ($n = 66$) for module 3. The repetitive and restricted (RRB) domain had ICCs of .79, .80 and .91 for modules 1, 2 and 3, while the ICCs of the overall totals were .97, .96 and .94, respectively. The interrater reliability of classification was reported in PA for modules 1–3 from still another subsample of the same dataset: 95 % for

**Table 1** Sample description

| | ASD | Non-ASD |
|---|---|---|
| *Module 1* | | |
| *N* (males, females) | 8 (7, 1) | 2 (1, 1) |
| Chronological age (years) (SD) | 3.77 (1.03) | 3.71 (.41) |
| Verbal IQ (SD) | 60.3 (19.6) | 77.5 (6.4) |
| Nonverbal IQ (SD)[a] | 88.0 (−) | 77.0 (−) |
| *Module 2* | | |
| *N* (males, females) | 8 (6, 2) | 2 (2, 0) |
| Chronological age (years) (SD) | 4.76 (.73) | 4.17 (.94) |
| Verbal IQ (SD) | 89.3 (15.6) | 94.0 (−) |
| Nonverbal IQ (SD) | – | – |
| *Module 3* | | |
| *N* (males, females) | 7 (6, 1) | 3 (3, 0) |
| Chronological age (years) (SD) | 11.12 (2.09) | 9.39 (2.80) |
| Verbal IQ (SD) | 92.3 (12.0) | 93.0 (0) |
| Nonverbal IQ (SD) | 102.5 (16.4) | 98 (-) |
| *Module 4* | | |
| *N* (males, females) | 5 (4, 1) | 5 (4, 1) |
| Chronological age (years) (SD) | 16.05 (1.19) | 15.98 (2.46) |
| Verbal IQ (SD) | 97.0 (15.2) | 104.0 (12.7) |
| Nonverbal IQ (SD) | 94.2 (9.9) | 101.4 (14.2) |

*ASD* autism spectrum disorder, *SD* standard deviation, *IQ* intelligence quotient, The ASD group comprised individuals with autistic disorder ($n = 13$), Asperger's disorder ($n = 6$) and PDD-NOS ($n = 9$) with and without comorbidities. In the non-ASD group, $n = 8$ had ADHD, $n = 1$ language disorder, $n = 1$ intellectual delay and $n = 2$ no diagnosis. IQ data were not available for all participants

[a] Only one participant, therefore no SD

module 1 ($n = 46$ autism; $n = 13$ non-spectrum), 98 % for module 2 ($n = 28$ autism; $n = 6$ non-spectrum) and 92 % ($n = 46$ autism; $n = 1$ autism spectrum). All 12 contributing examiners were thoroughly trained and had reached research reliability. They also attended weekly coding meetings during the study period with continuous checks of the interrater reliability on item level, see Online resource—Table 1 for further details. Aside from the ADOS authors, Bölte and Poustka [15] published interrater reliability data collected within the framework of an ASD genetics research project. Twelve individuals with autistic disorder (three for each module) were independently assessed by five raters. The interrater reliability of classification was $\kappa_w = 1.00$. In the context of two neuroimaging studies of adults with ASD, Bastiaansen and colleagues [20] examined the interrater reliability of ADOS module 4 in a mixed sample of $n = 38$ with high-functioning ASD, $n = 18$ with schizophrenia, $n = 16$ with psychopathy and $n = 21$ with typical development ($N = 93$) rated by five research reliable psychologists including two certified ADOS trainers [30]. The group of examiners attended to regularly group meetings to calibrate their ratings. Two examiners coded each assessment, one unaware of the clinical diagnosis of the participant. Mean $\kappa_w$ of the 21 included items was .66, with $\kappa_w > .60$ for 14 of the 21 items and none $\kappa_w < .50$. Most of the items of section D (RRB) and E (other abnormal behaviors) were excluded due to too few ratings other than zero. The interrater reliability for domain totals ranged from ICC = .79 (communication) to .92 (reciprocal social interaction as well as the overall total). The level of interrater agreement was $\kappa = .73$ (PA = 89.2 %) using the lower autism spectrum cutoff and dividing the sample dichotomously in ASD/non-ASD groups.

Despite the extensive findings on the ADOS interrater reliability in research settings mentioned above, to the authors' best knowledge there are very few published results on the interrater reliability of the ADOS among basically ADOS trained clinicians using the ADOS in daily clinical practice. One study reported interrater reliability data on domain totals involving less ADOS experienced clinicians. Nevertheless, the focus of the study was rather comparing the agreement between less experienced clinicians and research reliable ADOS experts (ICC = .79–.82) [31] than establishing interrater reliability on the ADOS among ordinary clinicians. The lack of clinical studies on ADOS objectivity is surprising and unfortunate, as the vast majority of the ADOS' administrations take place in clinical practice and because findings from the ADOS highly influence diagnostic decision making in many clinical settings. Moreover, objectivity findings from research settings might not easily compare to common clinical settings, eventually overestimating agreement in standard care settings, and be associated with a higher degree of misclassification. Because some scientific findings build on clinical practice (e.g., those from register-based studies) [32–34], the lack of data on ADOS objectivity in the clinic carries even an unknown risk of bias in research. Therefore, the present study investigated the spontaneous interrater reliability on item level, for domain totals and classification of the ADOS across various naturalistic clinical settings among clinicians with different levels of clinical experience and expertise using the ADOS.

## Methods

### Participants

Forty children and adolescents, each 10 examined with the ADOS modules 1, 2, 3 or 4, were included in the study, most of whom were males with ASD (see Table 1 for sample characteristics). All participants had been videotaped as part of regular clinical routine diagnostic evaluation at an outpatient clinic or specialized neuropsychiatric unit between 2011 and 2014. The study was approved by the Regional Board of Ethical Vetting, Stockholm, and

informed consent by the participants or their caregiver was collected.

## Procedure

Each videotaped ADOS administration was rated by five raters, i.e., 50 ratings for each module. Four of five raters were blind for diagnostic status of the individual examined with the ADOS, while the fifth rater was the clinician who originally had examined the participant in clinical practice. Fourteen psychologists and one pediatrician from 13 different clinical centers participated in the study. Thirteen clinicians rated between 1 and 8 administrations of modules 1 and 2 as well as between 1 and 7 administrations in module 3. Moreover, 15 clinicians rated between 1 and 6 administrations of module 4. All raters had attended ADOS basic clinical training, but the expertise and experience of using the ADOS varied substantially. Basic clinical training in Sweden consists of a 2½-day-long workshop on the ADOS's theoretical background, its principles of use, demonstration of all modules, and coding and discussion of individual administrations. No formal reliability checks are conducted. Three raters (SB, KO and EZ) were research reliable and certified ADOS trainers. Merely limited calibration in the form of two rater meetings prior to the study was scheduled, as the intention was to investigate the spontaneous or "true" interrater reliability of the ADOS in everyday clinical use.

## Measures

### *ADOS*

The ADOS is a standardized direct observation scale designed to capture important social–communicative behaviors as well as any stereotypic and repetitive behavioral features. These aspects are coded, typically from 0 (denotes no abnormality related to autism/as specified) to 2 (definite evidence of abnormality) and sometimes 3 (profound severity), in sets of items where a selection is combined to form totals used for the instrument's diagnostic algorithms.

The ADOS-2 consists of five different modules and eight algorithms depending on the individual's expressive language level and/or age in order to minimize the influence of expressive language and developmental level/age on the diagnostic evaluation. In this study, modules 1–4 of the ADOS-2 and their six algorithms were included. The totals and the classifications for modules 1–3 of the ADOS-2 were applied [7] as well as those of the revised algorithm for module 4 [22]. The ADOS-2 toddler module was not included in this study.

Each module consists of 29–34 items. A selection of the most diagnostically informative 14 items is combined and summed up to form the ADOS-2 diagnostic algorithms for modules 1–3, and 15 items in the revised algorithm of module 4. The algorithm totals are compared against diagnostic classification cutoffs for the ADOS/-2 classifications of autism and autism spectrum. An ADOS/-2 classification is not necessarily equivalent to a clinical diagnosis. Only the lower diagnostic threshold autism spectrum classification cutoff of the ADOS was considered in this study, i.e., all individuals with either autistic disorder, Asperger's disorder or PDD-NOS as a group were tested against the autism spectrum cutoff.

## Statistics

The $G(q,k)$ estimator, calculated with the SAS macro $G(q,k)$ provided by Putka et al. [35], was used to analyze the interrater reliability of items and totals. As our study design was neither fully crossed nor nested (i.e., ill-structured measurement design, ISMD) [35], the necessary assumptions for the most commonly used statistical methods like $\kappa_w$ for multiple users and intraclass correlation (ICC) were not fulfilled or their applicability incompletely described for the current design [35–37]. The $G(q,k)$ estimator used here has been described as a modified ICC (1, $k$). It estimates the rater main effect separately from the rater–subject interaction in unbalanced designs [35, 38] to yield coefficients analogous to ICC [39]. It has been demonstrated to produce a more accurate estimate of interrater reliability than ICC in ISMDs, especially preventing from the risk of underestimating interrater reliability [35, 38]. Like in the Lord et al. studies [6, 7], scores of 3 were converted to 2 in the analyses, except for item A1 ("overall level of non-echoed spoken language") where the scores of 0–4 were kept. Scores indicating "not applicable" (7 and 8) were treated as missing values. Items that had fewer than three ratings other than zero were excluded from the analyses. The interrater reliability for the diagnostic classification was analyzed using Fleiss' $\kappa$ for multiple raters [40] and Cohen's $\kappa$ [41]. For items and diagnostic classification, the interrater agreement [42] was analyzed using (exact) agreement in percent (PA), i.e., the number of agreements divided by the total number of observations. For the interpretation of the clinical significance of the interrater reliability coefficients, we considered coefficients below .40 as poor, .40–.59 fair, .60–.74 good and above .75 excellent [43]. For PA, 70–79 % agreement was evaluated to be fair, 80–89 % good and above 90 % excellent [44].

**Table 2** ADOS module 1: $G(q,k)$ indicating interrater reliability and PA indicating interrater agreement

| Item | PA | $G(q,k)$ | Lord et al. 2012 | |
|---|---|---|---|---|
| | | | PA | $\kappa_w$ |
| *Language and communication* | | | | |
| Overall level of non-echoed language | 60 | .94 | 95 | .85 |
| Frequency of vocalizations directed to others | 73 | .83 | 97 | .92 |
| Intonation of vocalizations or verbalizations | 99 | .83 | 84 | .63 |
| Immediate echolalia | 83 | .83 | 96 | .90 |
| Stereotyped/Idiosyncratic use of words or phrases | – | – | 90 | .78 |
| Use of other's body to communicate | 58 | .82 | 94 | .84 |
| Pointing | 61 | .91 | 86 | .66 |
| Gestures | 53 | .84 | 90 | .78 |
| *Reciprocal social interaction* | | | | |
| Unusual eye contact | 82 | .85 | 100 | 1.00 |
| Responsive social smile | 52 | .69 | 92 | .83 |
| Facial expressions directed to others | 60 | .75 | 89 | .68 |
| Integration of gaze and other behaviors during social overtures | 67 | .85 | 91 | .78 |
| Shared enjoyment in interaction | 42 | .74 | 91 | .76 |
| Response to name | 78 | .93 | 88 | .75 |
| Requesting | 50 | .72 | 88 | .64 |
| Giving | 46 | .56 | 95 | .85 |
| Showing | 65 | .83 | 88 | .71 |
| Spontaneous initiation of joint attention | 53 | .76 | 98 | .96 |
| Response to joint attention | 70 | .91 | 100 | 1.00 |
| Quality of social overtures | 66 | .85 | 94 | .85 |
| *Play* | | | | |
| Functional play with objects | 56 | .86 | 91 | .78 |
| Imagination/creativity | 69 | .80 | 90 | .73 |
| *Stereotyped behaviors and restricted interests* | | | | |
| Unusual sensory interest in play material/person | 62 | .75 | 81 | .57 |
| Hand and finger and other complex mannerisms | 51 | .48 | 93 | .83 |
| Self-injurious behavior | – | – | 98 | .97 |
| Unusually repetitive interests or stereotyped behaviors | 42 | .71 | 84 | .55 |
| *Other abnormal behaviors* | | | | |
| Overactivity | 57 | .89 | 91 | .71 |
| Tantrums, aggression, negative or disruptive behavior | 71 | .61 | 93 | .78 |
| Anxiety | 48 | .23 | 97 | .77 |

$G$ $(q,k)$ is a coefficient analogous to intraclass correlation and is derived from body of G theory. It is described in Putka and colleagues, 2008. *PA* percent exact agreement. The $\kappa_w$ (Cohen's weighted kappa) from the Lord studies (1999, 2000, 2012) are reported for comparison

## Results

### Interrater reliability and exact agreement for items

The interrater reliability of the individual items for all modules is presented in Table 2, 3, 4 and 5. For module 1, the median of the interrater reliability for all items was $G(q,k) = .83$, range .23–.94. All items except for three (one RRB algorithm item) exceeded $G(q,k) = .60$. The median of the PA for all items was 60 %, range 42–99 %. For module 2, the median was $G(q,k) = .74$, range .38–.91. All items except for five (three algorithm items including one RRB item) had an interrater reliability of $G(q,k) \geq .60$. PA was 65 % (median), range 40–80 %. The interrater reliability of module 3 was $G(q,k) = .74$ (median), range .30–.89, and nine items (five algorithm items including three RRB items) fell below $G(q,k) = .60$. The median of PA was 61.5 %, range 40–90 %. For module 4, $G(q,k) = .75$ (median), range .29–.92. Three items fell below .60 of which one algorithm item PA was 59.5 % (median), range: 49–84 %.

**Table 3** ADOS module 2: $G(q,k)$ indicating interrater reliability and PA indicating interrater agreement

| *Item* | PA | $G(q, k)$ | Lord et al. 2012 | |
|---|---|---|---|---|
| | | | PA | $\kappa_w$ |
| ***Language and communication*** | | | | |
| Overall level of non-echoed language | 80 | .62 | 96 | .89 |
| Speech abnormalities associated with autism (intonation/volume/rhythm/rate) | 60 | .74 | 98 | .93 |
| Immediate echolalia | 65 | .68 | 94 | .81 |
| Stereotyped/Idiosyncratic use of words or phrases | 70 | .91 | 85 | .61 |
| Conversation | 70 | .81 | 86 | .53 |
| Pointing | 60 | .85 | 85 | .56 |
| Descriptive, conventional, instrumental or informational gestures | 55 | .70 | 91 | .79 |
| *Reciprocal social interaction* | | | | |
| Unusual eye contact | 80 | .84 | 93 | .85 |
| Facial expressions directed to others | 65 | .70 | 80 | .45 |
| Shared enjoyment in interaction | 60 | .78 | 78 | .38 |
| Response to name | 76 | .77 | 84 | .52 |
| Showing | 70 | .71 | 83 | .59 |
| Spontaneous initiation of joint attention | 50 | .49 | 85 | .58 |
| Response to joint attention | 78 | .82 | 96 | .83 |
| Quality of social overtures | 50 | .51 | 89 | .71 |
| Amount of social overtures/maintenance of attention | 55 | .38 | 82 | .55 |
| Quality of social response | 50 | .74 | 91 | .70 |
| Amount of reciprocal social communication | 50 | .67 | 93 | .81 |
| Overall quality of rapport | 65 | .82 | 83 | .51 |
| *Play* | | | | |
| Functional play with objects | 40 | – | 98 | .89 |
| Imagination/creativity | 65 | .86 | 83 | .53 |
| *Stereotyped behaviors and restricted interests* | | | | |
| Unusual sensory interest in play material/person | 80 | .74 | 83 | .49 |
| Hand and finger and other complex mannerisms | 55 | .83 | 93 | .69 |
| Self-injurious behavior | – | – | 97 | .55 |
| Unusually repetitive interests or stereotyped behaviors | 50 | .48 | 93 | .48 |
| *Other abnormal behaviors* | | | | |
| Overactivity | 65 | .81 | 91 | .76 |
| Tantrums, aggression, negative or disruptive behavior | 80 | .43 | 94 | .75 |
| Anxiety | 80 | .82 | 96 | .78 |

$G(q,k)$ is a coefficient analogous to intraclass correlation and is derived from body of G theory. It is described in Putka and colleagues, 2008. *PA* percent exact agreement. The $\kappa_w$ (Cohen's weighted kappa) from the Lord studies (1999, 2000, 2012) are reported for comparison

## Interrater reliability of domain scores and ADOS-2 classification

The interrater reliability of domains and overall totals for the algorithms of the ADOS-2 [7] and the revised algorithm of module 4 [22] are presented in Table 6. For SA, the $G(q,k)$ ranged from = .86 to .92, for RRB from .45 to .90 and for the overall total score from .85 to .92.

The interrater reliability and interrater agreement of ADOS-2 classification, i.e., whether raters were consistent if the autism spectrum cutoff was met or not, were assessed with Fleiss' $\kappa$, Cohen's $\kappa$ and PA; results are presented in Table 7. Fleiss' $\kappa$ was .38 (range .19–.55) and Cohen's $\kappa$ was .69 (range .61– .76) for all modules 1–4 together. PA was 74.5 % for all modules (range 64–82 %) for modules 1–4. Agreement was associated with different ASD diagnoses. Autistic disorder was agreed on in 10 of 13 cases and in four of six cases of Asperger's disorder, while this was the case only for a minority of the individuals with PDD-NOS (two of nine). Only a minority of the participants with attention deficit/hyperactivity

**Table 4** ADOS module 3: $G(q,k)$ indicating interrater reliability and PA indicating interrater agreement

| Item | PA | $G(q, k)$ | Lord et al. 2012 | |
|---|---|---|---|---|
| | | | *PA* | $\kappa_w$ |
| *Language and communication* | | | | |
| Overall level of non-echoed language | 88 | .72 | 88 | .49 |
| Speech abnormalities associated with autism (intonation/volume/rhythm/rate) | 46 | .36 | 88 | .53 |
| Immediate echolalia | – | – | 92 | .69 |
| Stereotyped/Idiosyncratic use of words or phrases | 50 | .37 | 77 | .50 |
| Offers information | 47 | .79 | 85 | .64 |
| Asks for information | 63 | .89 | 83 | .50 |
| Reporting of events | 50 | .64 | 77 | .50 |
| Conversation | 40 | .66 | 87 | .68 |
| Descriptive, conventional, instrumental or informational gestures | 54 | .75 | 85 | .52 |
| *Reciprocal social interaction* | | | | |
| Unusual eye contact | 76 | .83 | 100 | 1.00 |
| Facial expressions directed to others | 49 | .83 | 88 | .68 |
| Language production and linked nonverbal communication | 60 | .79 | 92 | .81 |
| Shared enjoyment in interaction | 52 | .77 | 90 | .66 |
| Empathy/comments on other's emotions | 52 | .49 | 83 | .54 |
| Insight | 69 | .80 | 92 | .76 |
| Quality of social overtures | 60 | .40 | 87 | .61 |
| Quality of social response | 64 | .81 | 88 | .60 |
| Amount of reciprocal social communication | 44 | .66 | 85 | .62 |
| Overall quality of rapport | 57 | .78 | 88 | .68 |
| *Imagination* | | | | |
| Imagination/creativity | 70 | .85 | 85 | .54 |
| *Stereotyped behaviors and restricted interests* | | | | |
| Unusual in play material/person | 86 | .44 | 86 | .46 |
| Hand and finger and other sensory interest complex mannerisms | 72 | .53 | 90 | .47 |
| Self-injurious behavior | – | – | 100 | – |
| Excessive interest in or references to unusual or highly specific topics or objects or repetitive behaviors | 82 | .35 | 98 | .94 |
| Compulsions or rituals | 80 | .30 | 98 | .85 |
| *Other abnormal behaviors* | | | | |
| Overactivity | 74 | .87 | 81 | .60 |
| Tantrums, aggression, negative or disruptive behavior | 90 | .83 | 96 | – |
| Anxiety | 90 | .44 | 88 | .61 |

$G(q,k)$ is a coefficient analogous to Intraclass Correlation and is derived from body of G theory. It is described in Putka and colleagues, 2008. *PA* percent exact agreement. The $\kappa_w$ (Cohen's weighted kappa) from the Lord studies (1999, 2000, 2012) are reported for comparison

disorder [ADHD] (2 of 8) were categorized consistently across raters. When the participants with PDD-NOS and ADHD were removed from the analyses, Fleiss' and Cohen's $\kappa$ increased to .45 and .75, respectively, and PA to 86 %.

## Discussion

This study examined for the first time the interpersonal objectivity or interrater reliability of the ADOS in a naturalistic multicenter clinical setting. Overall, our results contribute to a better understanding of the psychometric properties of the ADOS in ordinary daily clinical use, while previous studies mostly reported interrater reliabilities between raters with thorough preparation and continuously and systematically updated calibration, such as in Lord [6, 7] and Bastiaansen [20]. This is significant information as the ADOS is widely used around the world in clinical practices, and recommended by different national guidelines, and professional societies, but clinicians

**Table 5** ADOS module 4: $G(q,k)$ indicating interrater reliability and PA indicating interrater

| Item | *PA* | $G(q, k)$ | Lord et al. 2012 | |
|---|---|---|---|---|
| | | | PA | $\kappa_w$ |
| *Language and communication* | | | | |
| Overall level of non-echoed language | – | – | 88 | – |
| Speech abnormalities associated with autism (intonation/volume/rhythm/rate) | 52 | .68 | 85 | .65 |
| Immediate echolalia | – | – | 95 | .64 |
| Stereotyped/idiosyncratic use of words or phrases | 62 | .65 | 91 | .66 |
| Offers information | 57 | .65 | 85 | .64 |
| Asks for information | 61 | .90 | 90 | .71 |
| Reporting of events | 50 | .29 | 85 | .65 |
| Conversation | 51 | .48 | 98 | .93 |
| Descriptive, conventional, instrumental or informational gestures | 56 | .64 | 80 | .50 |
| Emphatic or emotional gestures | 63 | .90 | 85 | .64 |
| *Reciprocal social interaction* | | | | |
| Unusual eye contact | 78 | .87 | 80 | .60 |
| Facial expressions directed to others | 49 | .68 | 93 | .72 |
| Language production and linked nonverbal communication | 81 | .91 | 85 | – |
| Shared enjoyment in interaction | 60 | .90 | 88 | .70 |
| Communication of own affects | 53 | .72 | 83 | .61 |
| Empathy/comments on other's emotions | 62 | .86 | 85 | .64 |
| Insight | 50 | .77 | 85 | .68 |
| Responsibility | 49 | .75 | 85 | .48 |
| Quality of social overtures | 64 | .84 | 88 | .69 |
| Quality of social response | 67 | .87 | 90 | .71 |
| Amount of reciprocal social communication | 57 | .85 | 85 | .65 |
| Overall quality of rapport | 59 | .78 | 93 | .79 |
| *Imagination* | | | | |
| Imagination/creativity | 71 | .92 | 83 | .57 |
| *Stereotyped behaviors and restricted interests* | | | | |
| Unusual sensory interest in play material/person | – | – | 98 | .84 |
| Hand and finger and other complex mannerisms | 84 | .65 | 91 | .66 |
| Self-injurious behavior | – | – | 100 | – |
| Excessive interest in or references to unusual or highly specific topics or objects or repetitive Behaviors | – | – | 85 | .41 |
| Compulsions or rituals | – | – | 90 | – |
| *Other abnormal behaviors* | | | | |
| Overactivity | 80 | .62 | 95 | .77 |
| Tantrums, aggression, negative or disruptive behavior | 96 | .62 | 90 | – |
| Anxiety | 64 | .37 | 90 | .68 |

$G(q,k)$ is a coefficient analogous to intraclass correlation and is derived from body of G theory. It is described in Putka and colleagues, 2008. *PA* percent exact agreement. The $\kappa_w$ (Cohen's weighted kappa) from the Lord studies (1999, 2000, 2012) are reported for comparison

are typically only basically trained on the ADOS, not "research reliable."

Findings indicate that the ADOS is a sufficiently objective measure even among non-calibrated clinicians with varying clinical experience from different clinical sites. For items and domain totals, the interrater reliability was basically in the same good to excellent range as reported in previous studies for extensively trained and calibrated research reliable raters, even though the exact agreement for items was substantially lower. Our sample was too small to analyze interrater reliabilities on item level across diagnoses. No consistent pattern of items with low and high interrater reliability across modules was detected in our sample, unless, possibly, a moderate tendency for the

**Table 6** $G(q,k)$ indicating interrater reliability for totals

| | Module 1 $n = 10$ | Module 2 $n = 10$ | Module 3 $n = 10$ | Module 4 $n = 10$ |
|---|---|---|---|---|
| Social affect | .91 | .86 | .86 | .92 |
| Repetitive and restricted behavior | .76 | .90 | .45 | .73 |
| Overall total | .92 | .89 | .85 | .90 |

RRB items like sensory interests, mannerisms, repetitive interests and compulsions and rituals to be less reliable than the SA items. This is consistent with previous studies that yielded generally lower interrater reliabilities of the RRB items compared with those of the SA. Reasons for the lower objectivity might be the often "low frequency" of RRB during the ADOS and therefore low coding threshold for RRB items, and that some clinicians might have divergent concepts of RRB and therefore score them on different ADOS RRB items. For example, the same behavior might be experienced by one clinician as a repetitive pattern, and as a compulsion, by another.

The objectivity for ADOS-2 classification was also in the acceptable range, although lower than in previous studies, and too low to support the notion that the ADOS might function as an ASD decision maker for non-expert examiners in everyday clinical practice or to replace best estimate consensus diagnoses. Using PA, only module 1 scored in the range of good objectivity, modules 3 and 4 in the fair range and module 2 in the poor range; all modules taken together yielded a fair objectivity for ADOS-2 classification. This study differs from the previous ones in several aspects, which might have contributed to the lower objectivity regarding ADOS-2 classification. Firstly, it is probable that the differences in ADOS training status, experience of using the ADOS and calibration efforts of the raters influenced the convergence of classification. Secondly, differences in sample characteristics might have contributed to the present results. A large proportion of the participants of the present study had lesser variants of ASD (PDD-NOS) or ADHD, an important differential and comorbid diagnosis of ASD, while the Lord et al. [6, 7, 26] and Bölte and Poustka [15] studies predominantly comprised individuals with core autistic disorder and only few individuals with milder ASD diagnoses and NS diagnoses (see Online resources Table 1 for details). When including participants with PDD-NOS, Lord et al. [6, 26] found lower objectivity in terms of PA, although still a good to excellent one (81–93 %). When excluding participants with PDD-NOS and ADHD in the present study, the PA increased from a fair (74.5 %) to a good level (86 %) (see Table 7). Thus, apparently, clinical heterogeneous samples, resembling clinical reality, are associated with less objective ADOS ratings and ADOS-2 classification, compared with more homogenous samples of ASD (resembling research settings), at least with non-expert raters. Clinicians should be well aware of that a consistently higher level of objectivity, even on categorical level, demands advanced training, and possibly ongoing calibration efforts, as well general experience of ASD, developmental psychology and psychopathology and psychometrics. Clinicians must also be aware of that the published validity data of the instrument have been generated by highly trained, research reliable examiners, and that presumably, the level of interrater reliability affects the validity. However, it is important to remind that these issues are surely not unique to the ADOS, but equally reflect the nature of the ASD concept and psychiatric diagnostic practice in general. For example, the interrater reliability of specific criteria of the DSM-IV was in the same range as for the different items of the ADOS ($\kappa = .58$–$.79$, PA = 82–93 %) [1] and for the classification of ASD versus non-ASD (DSM-IV: $\kappa = .95$; DSM-5: $\kappa = .69$) [45, 46]. Moreover, the interrater reliability for autistic disorder between experienced raters was higher than between inexperienced ones using or not using the DSM-IV criteria in their assessments ($\kappa = .84$ and $\kappa = .94$ vs. $\kappa = .34$ and $\kappa = .59$) and the interrater reliability increased for the inexperienced raters when using the DSM-IV criteria [1]. Therefore, in diagnostic decision making neither the ADOS nor any other current diagnostic instrument can replace and/or overrule "experienced clinical judgement" by an expert team using different pieces of systematically

**Table 7** Percent exact agreement (PA), Fleiss' generalized $\kappa$ and Cohen's $\kappa$ for the ADOS-2 classifications of autism spectrum versus non-spectrum

| | Module 1 | Module 2 | Module 3 | Module 4 | All | All PDD-NOS and ADHD excluded |
|---|---|---|---|---|---|---|
| PA (%) | 82 | 64 | 74 | 78 | 74.5 | 86 |
| Fleiss' $\kappa$ | .39 | .22 | .19 | .55 | .38 | .45 |
| Cohen's $\kappa$ | .71 | .61 | .62 | .76 | .69 | .75 |

*PDD-NOS* pervasive developmental disorder not otherwise specified, *ADHD* attention deficit/hyperactivity disorder, *PA* percent exact agreement. Module 1: $n = 5$ autistic disorder, $n = 3$ PDD-NOS and $n = 2$ non-ASD; module 2: $n = 4$ autistic disorder, $n = 4$ PDD-NOS and $n = 2$ non-ASD (all ADHD); module 3: $n = 3$ autistic disorder, $n = 4$ Asperger's disorder and $n = 3$ non-ASD (all ADHD); module 4: $n = 1$ autistic disorder, $n = 2$ Asperger's disorder, $n = 2$ PDD-NOS and $n = 5$ non-ASD ($n = 3$ ADHD)

collected information [47]. The latter is fully in line with statements put forward by the authors of ADOS concerning their recommendation of the use of the ADOS [7]; pp. 187, 208]. A too strong or naïve belief in results of diagnostic instrument like the ADOS might even potentially cause harm by erroneous diagnostic decision making [3].

The present study applied statistical methods that were different from those of previous studies to analyze the objectivity, because of our "unbalanced" design with multiple raters chosen to approximate clinical constellations. In such cases, the use of generalizability coefficients has been proposed as the method of first choice as it accounts for characteristics of the specific design [35, 38]. However, in the previous studies that most often used the same unbalanced design as ours, $\kappa_w$ was applied. The rationale for this choice of method is not explicitly described in these studies. Nevertheless, we assume that the choice was made in order to enable the use of different weights, i.e., linear instead of quadratic that takes into account the nature of ordinal-like scales containing a score of zero ("0"). As a consequence, the analyses in the previous studies using $\kappa_w$ might have generated conservative objectivity estimates compared to ours.

This study suffers from several limitations. First, for the present complex unbalanced design a larger and even more diverse sample would have been favorable to examine for example how certain participant characteristics and diagnoses might influence the different levels of ADOS objectivity. Second, this study did only examine objectivity, not diagnostic validity. As diagnostic validity is dependent on objectivity, good objectivity does not automatically translate into other properties such as diagnostic validity, solely analyzing the objectivity leaves us with incomplete psychometrics. However, as this is the first study investigating the objectivity of the ADOS among non-experts clinical users who form the majority of its users, it still adds novel evidence on the instrument's value in clinical practice.

In conclusion, we showed that (1) the objectivity or interrater reliability of the ADOS in clinical everyday settings among a variable group of mainly non-experts clinicians is good enough to warrant the use of it as a psychometrically sufficiently sound method to improve diagnostic decision making even in clinical settings and (2) the current state of evidence of the ADOS' psychometric properties does not favor its use as stand-alone diagnostic decision maker. Future research should focus on studies on the objectivity and diagnostic validity in large, heterogeneous but well-described samples of subjects and raters where several aspects are studied in the same sample to further elucidate the instrument's psychometric properties.

**Acknowledgments** We thank the Child and Adolescent Psychiatry of Stockholm County Council, Region Västra Götaland and Region Skåne, PRIMA Child and Adult Psychiatry, Tiohundra Norrtälje and Astrid Lindgren Children's Hospital and Sachs' Children and Youth Hospital, Stockholm County Council, for the generous cooperation in realizing this study. We also thank our funders Frimurare Barnhuset, Swedish Research Council, Stockholm County Council and finally Karolinska Institutet. Finally, we are especially grateful toward all the children and parents that participated in the project.

**Compliance with ethical standards**

**Conflicts of interest** On behalf of all authors, the corresponding author states that there is no conflict of interest.

## References

1. Klin A, Lang J, Cicchetti DV, Volkmar FR (2000) Brief report: interrater reliability of clinical diagnosis and DSM-IV criteria for autistic disorder: results of the DSM-IV autism field trial. J Autism Dev Disord 30(2):163–167
2. Volkmar F, Siegel M, Woodbury-Smith M, King B, McCracken J, State M, American Academy of C, Adolescent Psychiatry Committee on Quality I (2014) Practice parameter for the assessment and treatment of children and adolescents with autism spectrum disorder. J Am Acad Child Adolesc Psychiatry 53(2):237–257
3. Autism: Recognition, Referral, Diagnosis and Management of Adults on the Autism Spectrum (2012). The British Psychological Society & The Royal College of Psychiatrists. Leicester, UK
4. Swedish Council on Health Technology Assessment (SBU) (2013) Autismspektrumtillstånd. Diagnostik och insatser, vårdens organisation och patientens delaktighet—En systematisk litteraturöversikt, vol 215. Swedish Council on Health Technology Assessment
5. American Psychiatric Association (APA) (2013) Diagnostic and statistical manual of mental disorders DSM-5, 5th edn. American Psychiatric Association, Arlington
6. Lord C, Rutter M, DiLavore P, Risi S (1999) Autism Diagnostic Observation Schedule (ADOS). Western Psychological Publishing, Los Angeles
7. Lord C, Rutter M, DiLavore P, Risi S, Gotham K, Bishop S (2012) Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) Manual (Part I). Western Psychological Services, Torrance
8. Rutter M, Le Couteur A, Lord C (2003) Autism Diagnostic Interview Revised (ADI-R). Western Psychological Services, Los Angeles
9. Molloy CA, Murray DS, Akers R, Mitchell T, Manning-Courtney P (2011) Use of the Autism Diagnostic Observation Schedule (ADOS) in a clinical setting. Autism 15(2):143–162
10. Akshoomoff N, Corsello C, Schmidt H (2006) The role of the autism diagnostic observation schedule in the assessment of autism spectrum disorders in school and community settings. Calif School Psychol 11:7–19
11. Ashwood KL, Buitelaar J, Murphy D, Spooren W, Charman T (2015) European clinical network: autism spectrum disorder assessments and patient characterisation. Eur Child Adolesc Psychiatry 24(8):985–995
12. Wolff JJ, Gu H, Gerig G, Elison JT, Styner M, Gouttard S, Botteron KN, Dager SR, Dawson G, Estes AM, Evans AC, Hazlett HC, Kostopoulos P, McKinstry RC, Paterson SJ, Schultz RT, Zwaigenbaum L, Piven J (2012) Differences in white matter fiber tract development present from 6 to 24 months in infants with autism. Am J Psychiatry 169(6):589–600
13. Bryson SE, Zwaigenbaum L, Brian J, Roberts W, Szatmari P, Rombough V, McDermott C (2007) A prospective case series of

high-risk infants who developed autism. J Autism Dev Disord 37(1):12–24
14. Ozonoff S, Young GS, Belding A, Hill M, Hill A, Hutman T, Johnson S, Miller M, Rogers SJ, Schwichtenberg AJ, Steinfeld M, Iosif AM (2014) The broader autism phenotype in infancy: when does it emerge? J Am Acad Child Adolesc Psychiatry 53(4):398–407
15. Bölte S, Poustka F (2004) Diagnostic Observation Scale for Autistic Disorders: initial results of reliability and validity. Z Kinder Jugendpsychiatr Psychother 32(1):45–50
16. Zander E, Sturm H, Bölte S (2015) The added value of the combined use of the Autism Diagnostic Interview-Revised and the Autism Diagnostic Observation Schedule: Diagnostic validity in a clinical Swedish sample of toddlers and young preschoolers. Autism 19(2):187–199
17. Gotham K, Risi S, Pickles A, Lord C (2007) The Autism Diagnostic Observation Schedule: revised algorithms for improved diagnostic validity. J Autism Dev Disord 37(4):613–627
18. Gotham K, Risi S, Dawson G, Tager-Flusberg H, Joseph R, Carter A, Hepburn S, McMahon W, Rodier P, Hyman SL, Sigman M, Rogers S, Landa R, Spence MA, Osann K, Flodman P, Volkmar F, Hollander E, Buxbaum J, Pickles A, Lord C (2008) A replication of the Autism Diagnostic Observation Schedule (ADOS) revised algorithms. J Am Acad Child Adolesc Psychiatry 47(6):642–651
19. de Bildt A, Sytema S, van Lang ND, Minderaa RB, van Engeland H, de Jonge MV (2009) Evaluation of the ADOS revised algorithm: the applicability in 558 Dutch children and adolescents. J Autism Dev Disord 39(9):1350–1358
20. Bastiaansen JA, Meffert H, Hein S, Huizinga P, Ketelaars C, Pijnenborg M, Bartels A, Minderaa R, Keysers C, de Bildt A (2011) Diagnosing autism spectrum disorders in adults: the use of Autism Diagnostic Observation Schedule (ADOS) module 4. J Autism Dev Disord 41(9):1256–1266
21. Gray KM, Tonge BJ, Sweeney DJ (2008) Using the Autism Diagnostic Interview-Revised and the Autism Diagnostic Observation Schedule with young children with developmental delay: evaluating diagnostic validity. J Autism Dev Disord 38(4):657–667
22. Hus V, Lord C (2014) The autism diagnostic observation schedule, module 4: revised algorithm and standardized severity scores. J Autism Dev Disord 44(8):1996–2012
23. Kim SH, Lord C (2012) Combining information from multiple sources for the diagnosis of autism spectrum disorders for toddlers and young preschoolers from 12 to 47 months of age. J Child Psychol Psychiatry 53(2):143–151
24. Klein-Tasman BP, Risi S, Lord CE (2007) Effect of language and task demands on the diagnostic effectiveness of the autism diagnostic observation schedule: the impact of module choice. J Autism Dev Disord 37(7):1224–1234
25. 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. J Autism Dev Disord 38(2):362–372
26. Lord C, Risi S, Lambrecht L, Cook EH Jr, Leventhal BL, DiLavore PC, Pickles A, Rutter M (2000) The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. J Autism Dev Disord 30(3):205–223
27. Oosterling I, Roos S, de Bildt A, Rommelse N, de Jonge M, Visser J, Lappenschaar M, Swinkels S, van der Gaag RJ, Buitelaar J (2010) Improved diagnostic validity of the ADOS revised algorithms: a replication study in an independent sample. J Autism Dev Disord 40(6):689–703
28. 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. J Autism Dev Disord 19(2):185–212
29. DiLavore PC, Lord C, Rutter M (1995) The pre-linguistic autism diagnostic observation schedule. J Autism Dev Disord 25(4):355–379
30. de Bildt A, Sytema S, Meffert H, Bastiaansen JCJ (2015) The Autism Diagnostic observation schedule, module 4: application of the revised algorithms in an independent, well-defined, Dutch sample ($n = 93$). J Autism Dev Disord 30 August 2015. doi:10.1007/s10803-015-2532-4
31. McClure I, Mackay T, Mamdani H, McCaughey R (2010) A comparison of a specialist autism spectrum disorder assessment team with local assessment teams. Autism 14(6):589–603
32. Kim YS, Leventhal BL, Koh YJ, Fombonne E, Laska E, Lim EC, Cheon KA, Kim SJ, Kim YK, Lee H, Song DH, Grinker RR (2011) Prevalence of autism spectrum disorders in a total population sample. Am J Psychiatry 168(9):904–912
33. Fombonne E, Marcin C, Bruno R, Tinoco CM, Marquez CD (2012) Screening for autism in Mexico. Autism Res 5(3):180–189
34. Schendel DE, Diguiseppi C, Croen LA, Fallin MD, Reed PL, Schieve LA, Wiggins LD, Daniels J, Grether J, Levy SE, Miller L, Newschaffer C, Pinto-Martin J, Robinson C, Windham GC, Alexander A, Aylsworth AS, Bernal P, Bonner JD, Blaskey L, Bradley C, Collins J, Ferretti CJ, Farzadegan H, Giarelli E, Harvey M, Hepburn S, Herr M, Kaparich K, Landa R, Lee LC, Levenseller B, Meyerer S, Rahbar MH, Ratchford A, Reynolds A, Rosenberg S, Rusyniak J, Shapira SK, Smith K, Souders M, Thompson PA, Young L, Yeargin-Allsopp M (2012) The Study to Explore Early Development (SEED): a multisite epidemiologic study of autism by the Centers for Autism and Developmental Disabilities Research and Epidemiology (CADDRE) network. J Autism Dev Disord 42(10):2121–2140
35. Putka DJ, Le H, McCloy RA, Diaz T (2008) Ill-structured measurement designs in organizational research: implications for estimating interrater reliability. J Appl Psychol 93(5):959–981
36. Hallgren KA (2012) Computing inter-rater reliability for observational data: an overview and tutorial. Tutor Quant Methods Psychol 8(1):23–34
37. Conger AJ (1980) Integration and generalization of kappas for multiple raters. Psychol Bull 88(2):322–328
38. Cicchetti DV, Sparrow SA (1981) Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. Am J Ment Defic 86(2):127–137
39. Shavelson RJ, Webb NM (2006) Generalizability theory. Handbook of complementary methods in education research. Lawrence Erlbaum Associates Publishers, Mahwah, pp 309–322
40. Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychol Bull 76(5):378–382
41. Cohen J (1960) A Coefficient of Agreement for Nominal Scales. Educ Psychol Meas 20(1):37–46
42. de Vet HC, Terwee CB, Knol DL, Bouter LM (2006) When to use agreement versus reliability measures. J Clin Epidemiol 59(10):1033–1039
43. Cicchetti DV (1994) Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess 6(4):284–290
44. Cicchetti DV, Volkmar F, Klin A, Showalter D (1995) Diagnosing Autism using ICD-10 criteria: A comparison of neural networks and standard multivariate procedures. Child Neuropsychol 1(1):26–37
45. Volkmar FR, Klin A, Siegel B, Szatmari P, Lord C, Campbell M, Freeman BJ, Cicchetti DV, Rutter M, Kline W et al (1994) Field trial for autistic disorder in DSM-IV. Am J Psychiatry 151(9):1361–1367

46. Freedman R, Lewis DA, Michels R, Pine DS, Schultz SK, Tamminga CA, Gabbard GO, Gau SS, Javitt DC, Oquendo MA, Shrout PE, Vieta E, Yager J (2013) The initial field trials of DSM-5: new blooms and old thorns. Am J Psychiatry 170(1):1–5

47. Charman T, Gotham K (2013) Measurement Issues: Screening and diagnostic instruments for autism spectrum disorders—lessons from research and practice. Child Adolesc Ment Health 18(1):52–63