



ADOS-2 Module 4: Psychometric Properties and Diagnostic Performance at an Autism-specialized Clinic

Jens Christiansen¹ · Lennart Pedersen¹

Accepted: 9 July 2024

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024, corrected publication 2024

Abstract

Purpose Psychometric properties and diagnostic performance of the ADOS-2 module 4 were evaluated with participants from an autism-specialized clinic.

Methods The sample had 331 participants with 226 males and 70 females receiving an ASD diagnosis. The evaluation consisted of the following: (1) A replication of the Exploratory Factor Analysis (EFA) reported by Hus, V., & Lord, C. (2014). The Autism Diagnostic Observation Schedule, Module 4: Revised Algorithm and Standardized Severity Scores. *Journal of Autism and Developmental Disorders*, 44(8), 1996–2012. <https://doi.org/10.1007/s10803014-2080-3>. (2) Identification of ADOS-2 items best predicting clinical diagnosis using Recursive Feature Elimination (RFE) and comparison of these items to previous findings. (3) Receiver Operating Characteristic (ROC) curve analyses of the effects of age and IQ on diagnostic performance. (4) Comparisons of ADOS-2 revised algorithm scores between females and males and their association with ADI-R scores.

Results The EFA item-to-factor allocation of the ADOS-2 revised algorithm was largely reproduced. When comparing the present RFE to previous RFE findings, the item *Quality of Social Responses* stood out. ROC curve analysis showed outstanding diagnostic performance for adults with IQ above 70 but for females, ADOS-2 revised algorithm scores were lowered, and, contrary to males, did not correlate with ADI-R scores.

Conclusion Reproducing the item-to-factor allocation and finding outstanding agreement with the diagnostic decision for adults with IQ > 70 showcase the strength of the ADOS-2 revised algorithm. Furthermore, by incorporating, into future revisions, the finding of inter-clinic importance of the item *Quality of Social Responses*, performance might be further enhanced. Lastly, though, that female scores were lowered and did not correlate with ADI-R indicate a weakness in the ADOS-2 when applied to the adult female population.

Keywords ADOS · Module 4 · ASD · Autism Spectrum Disorder · Adult · Male · Female

Introduction

Autism spectrum disorder (ASD) is mainly characterized by deficits in initiating and sustaining social communication and reciprocal social interactions and the presence of restricted, repetitive, and inflexible patterns of behavior, interests, or activities (World Health Organization, 1992, 2019). Furthermore, ASD is a neurodevelopmental

condition, and its prevalence should therefore, logically, be constant across ages, which has also been confirmed epidemiologically (Brugha et al., 2011). Nonetheless, differences in the prevalence of ASD diagnoses by age indicate that there must be a very high number of undiagnosed adults. For example, in England, one study found that 1/23 of 10-19-year-old males have an ASD diagnosis but that the same can only be said for 1/550 of the 50-59-year-old males (O’Nions et al., 2023).

Being a neurodevelopmental condition, ASD traits must be present in early childhood. Sometimes, though, they are left unrecognized. This may happen for several reasons. One reason could be a lack of referral for specialist diagnostic assessment where the traits could have been described. Other reasons could be a misdiagnosis with a non-ASD

✉ Jens Christiansen
jens.h.christiansen@gmail.com

Lennart Pedersen
lennartpedersen@hotmail.com

¹ Center for Autism, Copenhagen, Denmark

condition having overlapping traits with those associated with ASD or that they had co-occurring diagnoses that made ASD traits harder to spot. Also, coping strategies, compensatory skills, and camouflaging behavior masking ASD traits might have developed over time, possibly since childhood (for more see Fusar-Poli et al., 2022).

Diagnostic assessment of any neurodevelopmental condition should include information about developmental history. Assessing adults for ASD presents itself with a specific challenge compared to assessing children. When assessing children for ASD an important source of information is the parent's recollections and descriptions of historical behavior. This source of information about childhood behavior and development will become less reliable over time because parents must try to remember behavior that happened decades ago and, sometimes for older adults, parents or other primary caregivers might not even be available for the diagnostic process (Lai & Baron-Cohen, 2015). Assessment of adults should therefore also rely on other sources of information.

One important source of information for identifying ASD in adults is self-report questionnaires. While it is important to include information from the person being assessed, self-report questionnaires present several shortcomings. A review of some of the most used self-report questionnaires concluded that they cannot stand alone for diagnostic purposes (Wigham et al., 2019). One example of a self-report questionnaire is the Autism-Spectrum Quotient (AQ) (Baron-Cohen et al. 2001). Ashwood et al. tested the performance of the AQ with informants who were participants consecutively referred to a national ASD diagnostic referral service for suspected ASD (Ashwood et al. 2016). All informants were assessed for possible ASD and also filled out the AQ. The specificity of the AQ was only 29% and consequently, the self-report version of the AQ could not predict the results of the diagnostic assessment.

The clinicians' observations must complement information about developmental history. Here the ADOS-2 module 4, designed for the assessment of verbally fluent adults/adolescents, usually at age 16 or more, is currently a widely accepted instrument used for this purpose. ADOS-2 module 4 offers a standardized semi-structured observational/conversational setting to assess communication, reciprocal social interaction, imagination/creativity, stereotyped behaviors, and restricted interests, to inform the diagnosis of ASD. ADOS-2 consists of five modules; a toddler Module, Modul 1, 2, 3, and 4, and can be used with individuals at a wide range of developmental and language levels (Lord et al., 2012). A prominent feature of module 4 is that the conversational interaction, between participant and clinician, is systematically

guided, in a standardized and structured way, thus helping the clinician apply a hierarchy of prompts and presses, to bring forth traits related to ASD (Lord et al., 2012). Following an ADOS-2 module 4 administration observed behavior is scored on 32 items (also counting item A1) and from these items, test developers have selected those most aligned statistically and conceptually with clinical diagnosis. These are referred to as diagnostic classification algorithms. Here our analytical focus will be on the performance of the newest algorithm, published in the scientific literature by one of the lead developers of the test, which we refer to as the ADOS-2 rev. alg. (Hus & Lord, 2014). Our analytic plan falls into four main parts. Below we expand on each of them.

Factor Structure of the ADOS-2 Revised Algorithm

According to DSM 5 ASD entails having two main traits: deficits in (1) reciprocal social interaction/social communication and the presence of (2) restricted and repetitive behavior. In the ADOS-2 rev. alg. module 4 10 items are selected to represent reciprocal social interaction/social communication (jointly called *social affect* (SA) and 5 items represent *restricted and repetitive behavior* (RRB) (Hus & Lord, 2014). Hus and Lord performed a two-factor EFA on the 15 items of the ADOS-2 rev. alg. which verified it and thereby validated their choice of items representing the two main traits of ASD. We will closely replicate the two-factor EFA performed by Hus and Lord. If we reproduce the factor structure then the instrumental properties of ADOS-2 module 4 and the ADOS-2 rev. alg. is further validated. If we cannot reproduce the factor structure it will be informative to know more about which items are loading on the factor opposite to the expected.

Which ADOS-2 Items Are Best at Predicting Clinical Diagnosis?

By examining which ADOS-2 rev. alg. items are most aligned with clinical diagnosis we can gain insight into the type of information clinicians on average put the highest weight on when deriving a clinical diagnostic decision. Küpper et al. (2020) examined which items were most aligned with clinical diagnosis using Recursive Feature Elimination (RFE). We want to compare the items that Küpper et al. found, to items found using RFE on the current data set. Identification of inter-clinical items best predicting diagnosis might, if present, provide valuable information about the relationship between the ADOS-2 rev. alg. and clinical diagnosis. Future revisions of the ADOS-2 rev. alg. might take advantage of evidence for items particularly related to clinical diagnosis.

Agreement Between Diagnostic Decision and the ADOS-2 Revised Algorithm Evaluated with Receiver Operating Characteristic Curves

Receiver Operating Characteristic (ROC) curves and computations of the Area Under the Curve (AUC) will be used to examine the agreement between the diagnostic decision and the ADOS-2 rev. alg. In the current data set the results of the ADOS-2 were allowed to influence the clinical diagnostic decision. This might cause what has been termed *diagnostic review bias* (Ransohoff & Feinstein, 1978). *Diagnostic-review bias* can cause both over- and underdiagnosis, because knowing the ADOS-2 rev. alg. score might influence how the clinicians interpret the rest of the diagnostic data, and this raises the possibility of overestimation of the accuracy estimates. Since we have no a priori reason to believe that the bias affects subgroups differently we try to address this issue by focusing on what ROC/AUC analysis reveals about relative differences between subgroups. Subgroups are: sex: male vs. female, age: child vs. adult, and, full-scale IQ: below vs. above 70) (see *Methods* for details on the definition of male/female).

We also compare the proposed ADOS-2 rev. alg. to the presently “official” ADOS-G algorithm. This comparison could prove informative for the future revision of the “official” algorithm in module 4 as well as for clinicians who still use the “official” algorithm (see *Methods*).

A companion instrument to ADOS-2 is the ADI-R which also measures autism by separating social and communicative traits and restricted and repetitive behavior (specifically the items from the diagnostic algorithm). Unlike the observational ADOS-2, the items from the ADI-R diagnostic algorithm use parents’ recollections, focusing on when the participant was between four and five years of age. Nevertheless, for both instruments, we can examine the relative contribution, to diagnostic decision, of measures of each of the two main ASD traits. Thereby we can gain information on whether the contributions are driven by the intrinsic properties of the instruments or if a contribution is also shaped by the emphasis the clinic places on a given trait, during the diagnostic procedure.

ADOS-2 Rev. Alg. Scores and Subgroups Based on Sex and Age

ADOS-2 rev. alg. scores: Differences between males and females: Does sex (assigned at birth) play a role for the ADOS-2 rev. alg. score in individuals with ASD? For children and adolescents (receiving modules 1–3), it was recently shown to be a very small effect (in an unprecedentedly large sample) (Kaat et al., 2021). In module 4 sex might be an important factor because some individuals,

apparently especially females, might be camouflaging, during the ADOS-2 administration, resulting in less visible ASD related traits (Hull et al., 2017). Furthermore, camouflaging could be a specific problem in module 4 because it might be present mostly in adults (Remnélius & Bölte, 2023). One way that camouflaging might work, with adult females in a module 4 context, could be related to stereotyped gender roles. For example, expressing emotions through gestures, language, or other means, during an ADOS-2 interview, might be enough to get a lower score on some ADOS-2 rev. alg. items (for example A10, B1, B2, and B5), and one well-known gender stereotype is that males do not express emotions as often as females. If it is found that females have lower ADOS-2 scores, then looking at measures of ASD traits, with an instrument not affected by camouflaging to the same extent, could be informative. If female scores are not lower than male scores, when measured on such an instrument, it lends support to the idea that camouflaging is reducing the visibility of ASD-related traits during the administration of the ADOS-2. Items from the diagnostic ADI-R algorithm could be less affected by camouflaging. For one, these items mostly relate to early childhood behavior where gender-stereotyped camouflaging probably is less developed (Remnélius & Bölte, 2023). Secondly, many items ask about behavior that is hard to camouflage over long periods of time (offering of comfort, lack of friends, routines/rituals, and highly focused prolonged interests). We will therefore compare ADI-R algorithm item scores between males and females and analyze the influence of sex on associations between ADI-R scores and ADOS-2 rev. alg. scores.

ADOS-2 rev. alg. scores: Differences between children and adults: Because the ADOS-2 is designed for adults, age might also play a role in the ADOS-2 rev. alg. score in individuals with ASD. We therefore looked at ADOS-2 rev. alg. scores in subgroups based on age being below/above 18 years using the same analyses as for males and females.

Methods

Participants

The full data set comprised 331 participants who all completed an ADOS. For some of these participants, it was possible to collect additional information from other measures. These are listed in Table 1. In virtually all ASD diagnostic evaluations at the autism specialized clinic (Center for Autism, Denmark) one of the five available ADOS-2 modules was used so there is no selection bias for using ADOS once referred to the clinic. For module 4 age and IQ were not limiting factors and only the clinician’s decision

Table 1 Sample characterization

Variables	Mean (SD)	Min-Max	N	Participants grouped by sex, age and diagnosis					
				Sex assigned at birth		Not ASD		Age	
				ASD M ⁹	F ¹⁰	M	F	ASD >	Not ASD <
Age	23 (8.3)	12–57	331	226	70	30	5	18	14
Year of diagnosis	2005 (4.9)	1995–2016							
ADOS-2 rev. alg. ¹									
Social affect (SA)	9.2 (4.6)	0–20							
RRB ²	2 (1.5)	0–9							
SA + RRB	11.2 (5.4)	0–27							
ADOS-G11 alg. ³	9.7 (4.6)	0–22							
ADI-R ⁴			296	204	64	24	4	205	10
Communication	9.5 (4.7)	0–22							
Social interaction	12.5 (6.3)	0–31							
RRSB ⁵	3.2 (2.3)	0–12							
FSIQ ⁶	91 (17)	45–141	259	177	56	22	4	183	10
VIQ ⁷	90 (17)	42–147	231	158	54	17	2	160	10
Vineland-II			122	85	28	9	0	86	2
Communication	64.6 (21.7)	20–109							
Socialization	60 (18.7)	20–111							
Daily living	75.4 (23.4)	20–126							
ABC ⁸	62 (16.3)	20–100							

¹Autism Diagnostic Observation Schedule 2 revised algorithm²Restricted, repetitive behavior³ADOS Generic⁴ADI-R: Autism diagnostic interview – revised⁵Restricted, repetitive, and stereotyped behaviors and interests⁶Full scale intelligence quotient: Wechsler type⁷Verbal intelligence quotient: Wechsler type⁸ABC: Adaptive Behavior Composite based on the 3 subscales.^{9/10}M=Male and F=Female

of the suitability of the module, for a specific individual, was considered (particularly that the individual was deemed verbally fluent). All referrals were referred because they were thought to have ASD. In a few instances, age was as low as 12 years and some individuals have a full-scale IQ as low as 45. These individuals might seem like a poor fit for Module 4 but, if deemed suitable for diagnostic use by the clinician, Module 4 can be used according to the earlier ADOS manual. Note, though, that the manual for ADOS-2 now recommends age to be ≥ 16 years (Lord et al., 2012). Most assessments were funded by the municipality to which the individual belonged. The overall mean age was 23 years for both sexes. Sex (male/female) is based on the Danish Central Person Registry (CPR). All 331 participants were recorded in the Central Person Registry with a personal identification number (CPR number). If the last four digits are an even/odd number it means that the participant was assigned the sex “female/male”, respectively, at birth.

Measures

Clinical Diagnosis During clinical procedures, several domains were systematically covered. These always included an assessment of core ASD traits, intelligence, language, and adaptive behavior (but not always using the same instruments). When deemed necessary, assessment of executive functions, attentional problems, or problems related to depression or anxiety was also part of the clinical procedure. The autism-specialized clinic was not a part of the national health system’s general psychiatry and broader psychiatric assessments were not performed. The autism-specialized clinic made systematic attempts at getting earlier records and, if available, medical history and previous testing were reviewed and included in the assessment. Individuals and, when possible, parents were interviewed by a clinician specialized in diagnosing ASD. For the assessment of ASD traits ADOS and ADI-R were used. Importantly, diagnostic decisions were never based on research algorithms of ADOS and ADI-R but always included contemplation of all item scores. All ADI-R and ADOS administrations were done by licensed clinicians with extended training in the clinical use of ADI-R/ADOS following the recommendations from the authors of the ADI-R/ADOS and supervised by on-site research reliable ADI-R/ADOS supervisors. During the period of data collecting the total group consisted of 9 clinicians/psychologists. All 9 clinicians had their reliability confirmed every sixth month by an on-site authorized ADI-R/ADOS supervisor. Intelligence was most often measured with a Wechsler-type test and assessment of adaptive behavior with Vineland-II Adaptive Behavior Scales. ADOS administration and scoring were always performed after the planning of the full diagnostic procedure

and simultaneously with the execution of the other diagnostic procedures, meaning that the choice of administering an ADOS was not biased by results from other tests and that the ADOS examiner and rater were blinded to all information about the individual that was not a necessary part of the ADOS process.

ADOS and ADOS-2 Both the Autism Diagnostic Observation Schedule and the Autism Diagnostic Observation Schedule, 2nd edition (ADOS-2) were used (Lord et al., 1999, 2012). According to the publishers, ADOS-2 is functionally identical to ADOS with the same activities and scoring system. For module 4 the difference between ADOS and ADOS-2 is the addition of a new non-algorithm item (“Amount of social overtures”). For brevity, we refer to both observation schedules as ADOS-2 throughout. In the current study, we will only analyze the 31 items shared by both ADOS versions. *The ADOS-2 rev. alg.*: According to the ADOS-2 rev. alg. an ASD threshold is reached when the sum of its 15 ADOS-2 items is at or above 8 (the higher the score the more autism traits). *The ADOS-G algorithm*: In select analyses, we will compare the ADOS-2 rev. alg. to the, for the time being, “official” module 4 algorithm from the most recent ADOS-2 Manual, published by Western Psychological Services (Lord et al., 2012). We will refer to this algorithm as the ADOS-G (Lord et al., 2000). This comparison is made to ensure that the diagnostic value of the proposed new set of items maintains at least the same level of accuracy. The ADOS-G algorithm is not just one number. Instead, an ASD threshold is reached when the sum of communication and social behavior domains are ≥ 7 , communication is ≥ 2 and social behavior is ≥ 4 . We will look selectively at the sum of the communication and social behavior domains of ADOS-G for analytical comparison with the ADOS-2 rev. alg. The sum of the communication and social behavior domains in ADOS-G is based on 11 items and therefore we will refer to it as ADOS-G11. When comparing the ADOS-G11 and the ADOS-2 rev. algorithms we are comparing the diagnostic value of items in the two algorithms.

ADI-R The autism diagnostic interview revised (ADI-R) is a manualized semi-structured lengthy interview with 93 items that, amongst others, systematically acquires historical early childhood data (Lord et al., 1994). Therefore, it heavily relies on information from parents. In the current dataset, all informants were parents. For individuals assessed before 2003, the ADI was used (LeCouteur et al., 1989). For brevity, we refer to both measures as ADI-R. For the current dataset, we extracted the ADI-R diagnostic algorithm total scores from the subscales of *Qualitative abnormalities in reciprocal social interaction*, *Communication*, and

Restricted, repetitive, and stereotyped patterns of behavior (Lord et al., 1994). Together these consist of 37 items and the range of possible summed scores is 0–68. The higher the score the more autism traits.

WAIS-R/WAIS-III Intelligence was measured using Wechsler Adult Intelligence Scale-Third Edition (WAIS-III; Wechsler, 1997) or Wechsler Adult Intelligence Scale-Revised (WAIS-R; Wechsler, 1981). In the current dataset, we used full-scale intelligence quotient (FSIQ) and verbal intelligence quotient (VIQ). In a few instances, the FSIQ was recently measured outside the clinic and in these cases, this FSIQ measure was used. In some of these cases, the VIQ was not available which resulted in a few more FSIQs than VIQs.

Vineland & Vineland-II The Vineland and Vineland-II Adaptive Behavior Scales (the survey form/parent interview) provides standardized index scores for three domains focusing on communication, daily living skills, and socialization and for a composite score of the three domains called the Adaptive Behavior Composite index score (ABC) (Sparrow et al., 1984, 2005). The three domains and the composite score do not change between versions and here, for simplicity, we refer to both as Vineland-II. Like the FSIQ, all index scores have a mean of 100 and a standard deviation of 15. The higher the scores the better the adaptive behavior. Between 1995 and 2010 the clinic used Vineland and Vineland-II to measure adaptive behavior. From 2011 to 2016 the clinic switched to other measures of adaptive behavior. These other measures of adaptive behavior are not available in the current dataset.

Procedure

Measures from main assessment instruments were collected and entered de-identified into the current dataset. The dataset consisted of the following variables: ASD decision, age, sex, year of assessment, ADOS-2 scores (scores for 31 individual items) ($n=331$), ADI-R score (algorithm domain scores only) ($n=296$), IQ score (Wechsler type FSIQ ($n=259$) and verbal scale ($n=231$) (domain scores only), Vineland-II score (domain scores only) ($n=122$). Vineland-II was included in the data set because it was the adaptive behavior instrument used most often. Because Vineland-II was only used 122 times we looked at the mean scores for FSIQ, ADI-R, ADOS-2 rev. alg. and age for this sub-group. These means were almost identical to the full sample.

Missing data: Out of the 31 ADOS-2 items in the current analysis, two items had a relatively large number of missing values. These were the items *Self Injurious Behavior*

and *Shared Enjoyment in Interaction* which had 22 and 93 missing values respectively. 18 of the missing *Self-Injurious Behavior* values and 89 of the *Shared Enjoyment in Interaction* values were missing because they were not included in an early prepublication version of the ADOS (Hus & Lord, 2014). These items were not part of the ADOS-G11 or the ADOS-2 rev. algorithms and were therefore kept in the dataset with missing as NA. For the remaining 29 items: 23 of the items had between zero and four missing values, and the remaining 6 items had between five and seven missing values (out of 331 possible missing pr. item). These were recoded into zeros in all analyses. According to the ADOS-2 manual, missing values are initially recorded with the number 9 and later converted to an algorithm score of 0 which is also the algorithm score for when there is no autism-specific behavior (Lord et al., 2012). Thus, advanced procedures for the handling of missing values are not a part of calculating an ADOS-2 module 4 algorithm score. The purpose of the present paper was to investigate the psychometric properties and diagnostic performance of the ADOS-2 module 4 in a real-life situation at an autism-specialized clinic. Therefore, advanced procedures, for handling missing data, were not applied, since they would thwart this purpose.

Data Analysis

Exploratory Factor Analysis (EFA) EFA was used to compare the latent factor structure of the current ADOS dataset to the published two-factor structure of the 15 ADOS items from the ADOS-2 rev. alg. (Hus & Lord, 2014). Like Hus and Lord (2014) we used participants with and without ASD ($n=331$). Factor analysis is often described as a large-sample procedure where the subject-to-variable ratio should be at least 10 (for example Norman & Streiner, 2014, p. 223). For this reason, EFA was not applied to subgroups such as male vs. female, age above or below 18 years, or IQ below or above an FSIQ 70. The EFA was made on the algorithm scores of either 0, 1, or 2. Because this data is ordinal Hus and Lord (2014) used polychoric correlations in the EFA. This was also the choice in the present analysis. The polychoric correlation matrix was derived with R using the function `polychoric()` in the `psych` package. The Kaiser-Meyer-Olkin test was used to test if the data was suited for factor analysis, by entering the polychoric correlation matrix into the function `KMO()` in the `EFStools` package[®]. For the 15 ADOS-2 rev. alg. items the KMO value was 0.827 which, according to Kaiser and Rice (Kaiser & Rice, 1974) means that the data is suitable for EFA. The purpose of the EFA was to see if the factor structure reported by Hus and Lord (2014) could be reproduced, and for this reason, the number of factors, to which the 15 items should be allocated, was set to two. The EFA was made in

R using the package Psych and the function fa() using the polychoric correlations. Factor method was principal axis. Like Hus and Lord (2014) we used Promax rotation. Using the items from the revised algorithm ADOS-2 module 4, Hus and Lord (2014) used confirmatory factor analysis (CFA) to evaluate model fit of item-to-factor allocation, for both a 1- and a 2-factor model. We replicated their analytical approach using cfa() from the R Lavaan package with ordered = TRUE and rotation = Promax.

Recursive feature analysis: RFE analysis was used to find the relative ADOS-item importance, among all items, for the prediction of clinical ASD decision. From 31 ADOS items Küpper et al. (2020) located the 5 most important in terms of predicting clinical ASD diagnosis. The 5 most important items reported by Küpper et al. were: (1) *Descriptive, Conventional, Instrumental, or Informational Gestures*, (2) *Unusual Eye Contact*, (3) *Facial Expressions directed to Others*, (4) *Quality of Social Response*, and (5) *Amount of Reciprocal Social Communication*. As can be

seen in Table 2 these five items are all part of the 11 items in the ADOS-G11 algorithm. Küpper et al. located the five items using RFE in R (rfe() from the Caret package with function = rfFuncs (random forest), method = repeatedcv, repeats = 5, and number = 10 (5-times repeated 10-fold cross-validation) and metric = kappa (hyperparameter tuning for set size of important items). We redid their analysis in R for all 331 individuals with the same functions and parameter settings (except repeats were set at 10 for a stable result of the first 5 items). Küpper et al. did the analysis on algorithm scores and, so did we. Küpper et al. imputed missing data and we recoded the few missing as zeros. To check for consistency of results between children vs. adults we also did a RFE analysis for these subgroups separately.

Receiver Operating Characteristic Curves The number of participants contributing to the ROC curves in Fig. 1 varies and is shown in Table 1.

Table 2 Items in ADOS-2 rev. and ADOS-G11 algorithms and results from RFE analysis

Items	ADOS-2 rev. Hus & Lord, 2014	ADOS-G11 Lord et al. 2000	Recursive feature elimination	
	Social affect	Social behavior + communication	Item importance rank	
			Current dataset ²	Küpper 2020
Conversation	X	X		
Emphatic or emotional gestures	X	X		
Unusual eye contact	X	X		2
Facial expressions directed at others	X	X		3
Quality of social overtures	X	X	1	
Quality of social responses	X	X	2	4
Amount of reciprocal social communication	X	X		5
Overall Quality of Rapport	X		3	
Communication of Own Affect	X			
Insight	X		4	
	Restricted and repetitive behavior			
Speech Abnormalities Associated with Autism	X ¹			
Unusual sensory interest	X			
Hand and finger mannerisms	X			
Exces. Inter. or ref. unus/high speci. topic/object	X			
Stereotyped/idiosyncratic use of words or phrases	X	X		
Responsibility		X		
Empathy/comments on others' emotions		X		
Descriptive, conventional, instrumental gestures		X		1

¹*Speech Abnormalities Associated with Autism* is the only item in the two-factor structure of the current dataset that did not follow the distribution of items (by highest loading) in the two-factor structure of Hus and Lord (2014). In the current dataset, this item loaded more on the social affect factor reported by Hus and Lord (2014)

²Fifth most important item in the current dataset was “*offers information*” (not shown because not part of ADOS-G11 or ADOS-2 rev. algorithms)

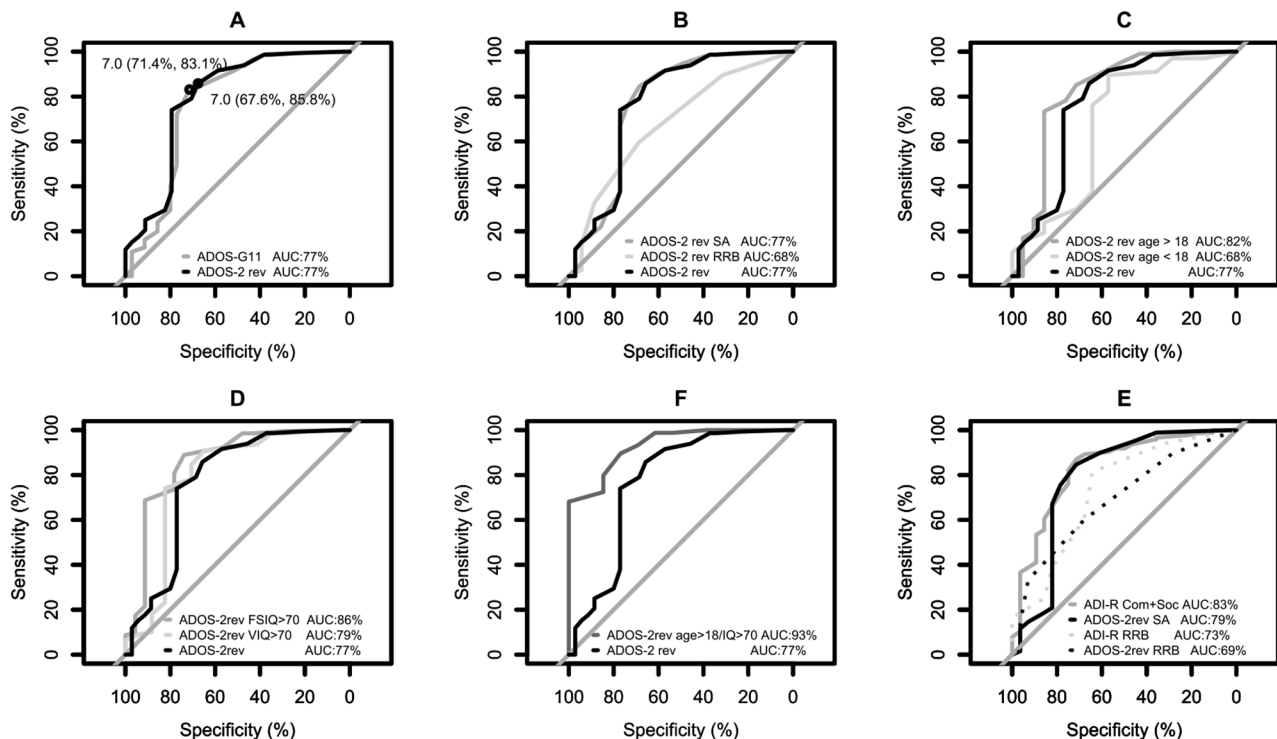


Fig. 1 Receiver Operating Characteristic curves (ROC curves): AUC: Area Under the Curve. ADOS-2 rev.: ADOS-2 revised algorithm (Hull and Lord 2014). ADOS-G11: Sum of the 11 items in the earlier ADOS-G algorithm. SA: social affect. RRB: restricted repetitive behavior. FSIQ: Full-scale IQ. VIQ: Verbal IQ. ADI-R com + soc: sum of the

two ADI-R domains communication and social interaction. ADI-R RRB: the ADI-R restricted repetitive behavior domain. Panel A: Circles on the two graphs are optimal cut-offs (7 for both) maximizing sensitivity and specificity (in parenthesis)) (full sample with $N=331$)

Table 3 Differences¹ between participants with ASD grouped by sex

Variables		Mean		t-test			
		Female	Male	t-value	df	p^2	CI
ADOS-2 rev. alg. SA + RRB	n=F70/M226	10.30	12.20	-2.90	119.00	0.016	-3.236 -0.612
FSIQ	n=F64/M204	91.00	90.60	0.15	107.00	0.884	-4.375 5.072
ADI-R sum of 3 domains	n=F56/M177	26.10	26.50	-0.24	124.00	0.884	-3.110 1.139
Vineland-II ABC	n=F28/M85	62.70	61.70	0.30	55.00	0.884	-5.574 7.544

¹Welch's t-test

²Hommel adjusted p -values. 4 comparisons.

ADOS-2 rev. alg. scores: Differences between male and female: To understand potential sex differences when using the ADOS-2 rev. alg. we will compare male/female differences in ADOS-2 rev. alg. means to differences in means of the sum of the three ADI-R domain scores, differences in means of the FSIQ, and differences in means of the Vineland-II ABC. For all tests of differences in means, we will be using Welch's t-test Hommel-corrected for 4 comparisons. Hommel correction was implemented in R. Furthermore, we will report Pearson's correlation between the SA and RRB factors of the ADOS-2 rev. alg. and the sum of the communication and social interaction domains of the ADI-R, the RRB domain of the ADI-R, the FSIQ, and the Vineland-II ABC. P -values of the correlations will be

Hommel-corrected for four comparisons. Results are shown in Tables 3 and 4.

ADOS-2 rev. alg. scores: Differences between children and adults: We replicated the above analyses but for children/adults (below/above 18 years). These are shown in Tables 5 and 6.

Results

Exploratory Factor Analysis

Except for one item (*speech abnormalities associated with autism*), the EFA reproduced the item-to-factor allocation of

Table 4 Correlations between ADOS-2 and other diagnostic measures: female/male participants with ASD

Variables			SA				RRB			
			r^1	CI ²	p^3		r^1	CI ²	p^3	
Female	ADI-R soc + com	$n = 64$	0.07	-0.20	0.30	0.689	0.01	-0.20	0.30	0.959
	ADI-R rrb	$n = 64$	-0.05	-0.30	0.20	0.689	0.11	-0.10	0.30	0.788
	FSIQ	$n = 56$	-0.26	-0.50	0.00	0.212	-0.26	-0.50	0.00	0.224
	Vineland-II ABC	$n = 28$	-0.08	-0.40	0.30	0.689	0.26	-0.10	0.60	0.564
Male	ADI-R soc + com	$n = 204$	0.35	0.20	0.50	< 0.000	0.16	0.00	0.30	0.069
	ADI-R rrb	$n = 204$	0.12	0.00	0.30	0.086	0.26	0.10	0.40	< 0.000
	FSIQ	$n = 177$	-0.39	-0.50	-0.30	< 0.000	-0.13	-0.30	0.00	0.078
	Vineland-II ABC	$n = 85$	-0.37	-0.50	-0.20	0.002	-0.20	-0.40	0.00	0.078

¹Pearson's correlation coefficient²95% confidence interval.³Hommel-adjusted p -values. 4 comparisons.Significant correlations with the lowest CI for r at or above weak correlation ($r=[0.2, 0.4]$) are in bold**Table 5** Differences¹ between participants with ASD grouped by age

Variables			Mean		t-test			
			< 18	=> 18	t-value	df	p^2	CI
ADOS-2 rev. alg. SA + RRB	$n = F67/M229$		11.60	11.80	-0.15	105.00	0.901	-1.513 1.294
FSIQ	$n = F50/M183$		26.80	26.20	0.40	121.00	0.901	-2.224 3.355
ADI-R sum of 3 domains	$n = F63/M205$		91.80	90.40	0.49	71.00	0.901	-4.380 7.233
Vineland-II ABC	$n = F27/M86$		61.60	62.10	-0.13	44.00	0.901	-7.706 6.803

¹Welch's t-test²Hommel adjusted p -values. 4 comparisons.**Table 6** Correlations between ADOS-2 and other diagnostic measures: above/below 18 years participants with ASD

Variables			SA				RRB			
			r^1	CI ²	p^3		r^1	CI ²	p^3	
Below 18	ADI-R soc + com	$n = 63$	0.41	0.20	0.60	0.003	0.13	-0.10	0.40	0.455
	ADI-R rrb	$n = 63$	0.15	-0.10	0.40	0.254	0.10	-0.20	0.30	0.455
	FSIQ	$n = 50$	-0.48	-0.70	-0.20	< 0.000	-0.16	-0.40	0.10	0.413
	Vineland-II ABC	$n = 27$	-0.49	-0.70	-0.10	0.009	-0.16	-0.50	0.20	0.413
Above 18	ADI-R soc + com	$n = 205$	0.26	0.10	0.40	< 0.000	0.13	0.00	0.30	0.136
	ADI-R rrb	$n = 205$	0.08	-0.10	0.20	0.24	0.28	0.10	0.40	< 0.000
	FSIQ	$n = 183$	-0.33	-0.50	-0.20	< 0.000	-0.16	-0.30	0.00	0.066
	Vineland-II ABC	$n = 86$	-0.24	-0.40	0.00	0.024	-0.10	-0.30	0.10	0.36

¹Pearson's correlation coefficient²95% confidence interval.³Hommel-adjusted p -values. 4 comparisons.Significant correlations with the lowest CI for r at or above weak correlation ($r=[0.2, 0.4]$) are in bold

the revised algorithm as reported by Hus and Lord (2014). Together, the two factors accounted for 43% of the variance with SA accounting for 34% and RRB accounting for 9%. The one item that differed, “*speech abnormalities associated with autism*”, had a higher loading on SA than on RRB. Root mean square error approximation (RMSEA) was 0.115 and the Tucker Lewis Index (TLI) was 0.794. The two-factor model was better than a one-factor model but the number of factors had to be raised to six to get RMSEA values under 0.08 and a TLI over 0.9. The Pearson's (product-moment) correlation between the factors SA and RRB

was 0.45. Compare this correlation to $r=0.46$ reported by Hus and Lord (2014). Hus and Lord reported a CFA with a CFI of 0.93 for a 2-factor solution and a CFI of 0.91 for a 1-factor solution (2014). They did not specify if the CFI test statistics were standard, scaled, or robust. Using the Multivariate Normality Test (mult.norm()) in the R package QuantPsyc to test the assumption of Multivariate Normality we found that both skewness and kurtosis had p -values far below 0.001 and therefore reject the hypothesis that our data follows a multivariate distribution. Followingly, Scaled and Robust test statistics are also reported: For a 2-factor model

with item-allocation to each factor as suggested by Hus and Lord (2014) the following was found: CFI: Standard=0.98, Scaled=0.95, Robust=0.81. 1-factor CFI: Standard=0.98, Scaled=0.94, Robust=0.8. For a 2-factor model with item-allocation to each factor as suggested by the EFA in the current study, where the item “speech abnormalities associated with autism” was placed in the SA factor instead of the RRB factor, the following was found: CFI: Standard=0.99, Scaled=0.96, Robust=0.87.

Recursive Feature Elimination

For all participants, the order of the importance of the five most important items was as follows: (1) *Quality of social overtures*, (2) *Quality of social response*, (3) *Overall quality of rapport*, (4) *Insight*, and (5) *Offers information*. These first five items were stable across 10 runs. In Table 2 the five most important items from the current dataset are listed together with the five most important items reported by Küpper et al. (2020). As can be seen in Table 2 4/5 of the most important items in the current dataset were from the SA domain and 0/5 from the RRB domain. Notice that 4/5 of the most important items that Küpper et al. reported also were from the SA domain (the only overlap between the two sets of five most important items is “*Quality of social responses*”) and 0/5 were from the RRB domain.

We also tried the RFE on two age groups (below/above 18 years). For participants below 18 years the five most important items were (1) *Insight*, (2) *Responsibility*, (3) *Overall quality of rapport*, (4) *Quality of social overtures*, and (5) *Reporting of events* (SA: Module 3 alg.). For participants above 18 years the five most important items were (1) *Quality of social overtures*, (2) *Empathy/comments on others' emotions*, (3) *Amount of reciprocal social communication*, (4) *Conversation*, and (5) *Quality of social response*. For both age groups, 4/5 top five items were from the SA or social/communication domain and no items were from the RRB domain. Furthermore, for both age groups and the full dataset *Quality of social overtures* was among the top five items. For the subgroup age > 18, the full dataset and in Küpper et al. the item *Quality of social response* was among the top five items.

Receiver Operating Characteristic Curves

From Fig. 1A it is evident that ROC curves of the ADOS-2 rev. and the ADOS-G11 algorithms behave similarly and both have an AUC of 77%. Hus and Lord (2014) choose the cut-off of 8, for the ADOS-2 rev. alg., by making a ROC analysis and selecting the cut-off that maximized sensitivity and specificity. As can be seen in Fig. 1A there is a dip in the graph for ADOS-2 rev. alg. after 7 and the sum of sensitivity

and specificity was marginally larger at both 7 and 9 than at 8. Nevertheless, the optimal cut-off for the current data is close to the official cut-off of 8. With a cut-off set at 8, for the full sample in 1A, the ADOS-2 rev. alg. had a specificity of 0.69 and a sensitivity of 0.79. The official ASD cut-off for the ADOS-G11 alg. is ≥ 7 and this was also the optimal cut-off with the current dataset. In Fig. 1B we see that SA alone is almost identical to SA+RRB (both AUC 77%) and that the RRB has a very low AUC. In Fig. 1C the 331 individuals are divided into groups based on age (≥ 18 ($n=250$) or < 18 ($n=81$)). This division reveals that the ADOS-2 rev. alg. works best with adults (AUC 82%) and not as well with adolescents aged 12–18 years (AUC 68%). In Fig. 1D we see that IQ also influences the performance of ADOS-2 rev. alg. There was only FSIQ data for 259 individuals and VIQ for 231 (see Table 1) so these ROC curves are based on subsets of the group used to draw the black ROC curve shown in Fig. 1A-E. Picking only individuals with an FSIQ ≥ 70 ($n=231$) increases the AUC to 86% and picking only individuals with a VIQ ≥ 70 ($n=206$) results in an AUC of 78%. When only looking at male individuals (not shown) there is no change in AUC (78%). We did not produce a ROC curve for females only since there were only 5/75 with no clinical ASD (see Table 1). We also looked at the effect of the year of diagnosis (not shown) and found no effect on the AUC. A follow-up Welsh t-test found no significant difference between the mean of SA+RRB score from assessments before 2006 (mean=11.52) and after 2006 (mean=10.85) (2006 is a median split of the year of diagnosis) ($t(327) = -1.1$, $p=0.26$). In Fig. 1E, we look at participants with FSIQ above 70 and age above 18 years ($n=176$). There is an increase in the AUC to 93%. Very few participants had FSIQ below 70 combined with being under 18 years of age so a ROC curve cannot be drawn for this subgroup. In Fig. 1F the ROC curve for the sum of the communication and social interaction domains of the ADI-R ($n=296$) has an AUC of 83% and the ROC curve for the RRB domain of the ADI-R has an AUC of 73%. When creating ROC curves for the corresponding SA and RRB domains of the ADOS-2 rev. alg., using the same 296 participants, The AUC for the SA ROC curve is 79% and the AUC for the RRB ROC curve is 69%.

Differences between Males and Females

There was no difference in diagnostic rate by sex in any analysis evaluating sex differences. In Table 3 we see that females score significantly lower (2 points) on the ADOS-2 rev. alg. For all 15 items, individually analyzed, the mean male score was higher than the mean female scores so, whatever caused the lower scores for females was a generalized factor. Nonetheless, the difference was higher for

some items. The two items with the largest difference in score, between sexes, were *Communication of own affect* followed by *Emphatic or emotional gestures*, accounting for about a third of the overall difference. In contrast, male and female scores are virtually identical for FSIQ, ADI-R, and Vineland-II ABC. Replacing VIQ with FSIQ produced similar results (not shown).

Table 4 shows, for males and females with ASD, correlations between SA or RRB of the ADOS-2 rev. and the sum of the two ADI-R domains *Social Interaction* and *Communication*, the ADI-R domain *Restricted Repetitive Behavior*, FSIQ, and Vineland-II ABC. Marked in bold are significant correlations where the lowest CI (95% Confidence Intervals) is at or above weak correlation (with weak defined as $r = (0.2, 0.4)$). No correlations for females are significant which is to be expected since very small correlations demand large sample sizes for determination of significance. For males, we see a different picture. Here we find significant borderline moderate correlations between SA and the sum of the two ADI-R domains *Social Interaction* and *Communication*, FSIQ and Vineland-II ABC. According to Fisher's R to Z transformation the correlations, for males and females with ASD, between ADI-R soc+com and ADOS-2 SA, were significantly different from each other ($z=2.02$, $p=0.043$ (two-tailed)). We will return to this important result in the discussion.

ADOS-2 Rev.Alg. Scores: Differences Between Children and Adults

A comparison of children and adult mean scores for ADOS-2 rev. alg., sum of three ADI-R domains FSIQ, and Vineland-II ABC, showed they were, clinically speaking, virtually identical, and no significant differences were observed (Table 5). In Table 6 we see that, for both age groups, the correlation between FSIQ and SA is significant and has the lowest CI value at or above weak correlation. The correlation is strongest for the group below 18 years and at $r = -0.48$ it is considerable. Also, in bold in Table 6, we see that the correlation between ADOS-2 SA and the two ADI-R domains *Social Interaction* and *Communication* is 0.41, for the group below 18 years, which can be compared to $r=0.26$ for the group above 18 years. There were no correlations between ADI-R RRB and ADOS-2 SA/RRB that were both significant and had CIs at or above weak correlation.

Discussion

The factor structure found by Hus and Lord (2014) was largely reproduced. In the current EFA one item loaded more on SA than on RRB relative to what was found by Hus

and Lord in 2014, namely “*Speech abnormalities associated with autism*”. It is an item that in an earlier version of ADOS was part of the communication domain but was not included in the ADOS-G algorithm and then later placed in the RRB domain in the ADOS-2 rev. alg. On closer examination, the item “*Speech Abnormalities Associated with Autism*” is characterized by having both communicative and restricted, repetitive, stereotyped, and inflexible elements. For this reason, it is perhaps not surprising that the item can switch places between factors SA and RRB. Future versions of the ADOS might benefit from replacing this item with a factor-wise, “*cleaner*” item. Nevertheless, the almost full replication of the factor structure, of the 15 items in the ADOS-2 rev. alg., does contribute to validating the proposed factors SA and RRB.

In both the current dataset and in Küpper et al. (2020), the RFE analysis showed that 4/5 items, best at predicting clinical diagnosis, were SA-items and that 0/5 were from the RRB domain. RFE analyses made separately for participants below/above 18 years gave the same result. This pattern of results is mirrored in the analyses with ROC curves in Fig. 1B where the ADOS-2 rev. SA and RRB scores have AUCs of 79% and 68% respectively. The interpretation of these findings is not straightforward, since, on the one hand, it could mean that the SA domain has superior instrumental properties relative to the RRB domain, but on the other hand, it could also mean that the present ASD specialized clinic, during the diagnostic procedure, places more emphasis on SA traits than on RRB traits. In Fig. 1F we observe the same pattern when drawing one ROC curve for the sum of ADI-R *Social interaction* and *Communication* domains (similar to the ADOS-2 SA domain) and another for the ADI-R *Restricted and Repetitive Behavior* domain (similar to the ADOS-2 RRB domain). The most straightforward explanation for these findings is that our ASD-specialized clinic, in arriving at a diagnostic decision, places most of the emphasis on SA traits.

The item *Quality of social responses* warrants a closer look because it was in the top five for the subgroup age > 18, for the full dataset, and for Küpper et al. (2020). This means it was in the top five for the age group most suited for ADOS-2 module 4 and across clinics. *Quality of social responses* is a summary item focusing broadly on the individual's social responses during the ADOS administration. A range of appropriate responses that are varied according to immediate social situations and presses will result in a score of 0, limited, socially awkward, inappropriate, or consistently negative is a score of 1, and odd, stereotyped responses that are restricted in range or inappropriate to the context will result in a score of 2, and minimal or no response to the examiner's attempts to engage the participant will result in a score of 3 (see Modul 4 protocol in

Lord et al., 2012). Thus, it seems that the presence/absence of suitably and contextually adapted reactions to the examiner's overtures are particularly important for diagnostic decisions. According to a recent study, the item *Quality of Social Responses* also has a particularly high value in terms of predicting yes/no to an ASD diagnosis for modules 2 and 3 so its importance might stand out across modules as well (Wolff et al., 2022).

In Table 3 we focus on score differences, on four main instruments, between participants with ASD grouped by sex. We found no differences in means between sexes for FSIQ, Vineland-II ABC, or ADI-R (and both male and female groups had a mean age of 23 years), but mean ADOS-2 rev. alg. score was two points larger for males. Interestingly, Fusar-Poli et al. (2022) reported a similar result, using what we here refer to as the ADOS-G11 algorithm, showing a mean ADOS-G11 score of 9 for males but only 7 for females. One possible explanation, for the differences between the sexes on the ADOS-2 rev. alg., is that males and females, having the same amount of ASD traits (according to ADI-R), and similar adaptive behavior and IQ, can differ on ADOS-2 because females are camouflaging during the ADOS-2 administration and thereby lowering their score on certain items. In our introduction, we suggested that gender-stereotyped behavior (for example that males do not express emotions as openly as females do) could drive the lower score for females. In line with this suggestion, we found that *Communication of own affect* and *Emphatic or emotional gestures* were the two items with the largest score difference between the sexes accounting for a third of the overall difference.

The two points lower ADOS-2 rev. score, for females with ASD, indicates that sex does play a role in the performance of the ADOS-2 rev. alg. Nevertheless, looking at the CIs in Table 3 we see that they are rather wide going all the way down to around a half-point difference between males and females for ADOS-2 rev. alg. But then again, from Table 4 we see that, for females, there is no correlation between the SA domain of the ADOS-2 rev. alg. and the sum of the *Social interaction* and *Communication* domains of the ADI-R (females: $r=0.07$). For males, this correlation is $r=0.35$. This finding adds to the worrying finding of a lower ADOS-2 rev. score for females, since it suggests that female ADOS-2 rev. scores are not only lower than male scores, but also that there is no association between them and scores from the ADI-R *Social interaction* and *Communication* domains.

The analyses of differences between means, when grouped into children and adults did not reveal any salient differences. It may be noted, though, that the correlation between ADOS-2 rev. and the sum of the ADI-R *Social interaction* and *Communication* domains was somewhat

larger for the group below 18 years, which, as mentioned in the introduction, can be due to the parent's recollections of historical behavior becoming less reliable over time. Also, with $r=-0.48$ (Table 6) the correlation between FSIQ and SA, for the group with ASD below 18 years, intelligence does seem to play a considerable role in the SA score.

Limitations

Gender vs. sex Assigned at Birth This study used sex assigned at birth to categorize participants. Sex assigned at birth is a biologically based binary category. This study did not have information on gender and our results do not capture gender differences.

Spectrum effect When considering the results of the ROC analyses, in the current paper, the reader should carefully consider if they apply to the context in which the reader wishes to use them. For example, in samples consisting of both participants thought to have ASD and participants not thought to have a psychiatric diagnosis, specificity can be higher relative to samples consisting of participants thought to have ASD and participants with other types of psychiatric diagnoses. The reason is that the two types of participants, in the first type of sample, will be easier, for the ADOS, to tell apart, than the two types of participants in the second sample, because the other types of psychiatric diagnoses can have features overlapping with ASD. Collectively, the factors that determine the characteristics of the sample, are referred to as *spectrum bias* or *spectrum effect* factors (Mulherin & Miller, 2002; Ransohoff & Feinstein, 1978). In the current sample, some of these *spectrum effect* factors were: (1) Only individuals referred for assessment were sampled, (2) the clinic is an autism-specialized center located outside a national health systems general psychiatry, and participants are referred for clarification of ASD status, (3) prevalence of ASD is high (89%) One systematic review on test accuracy in general advises clinicians to base diagnostic decisions on studies closely matching their clinical situation with prevalence in mind (Leeflang, 2008).

ADOS-2 rev. alg. scores and subgroups based on IQ Analyses of differences between ADOS-2 rev. alg. scores for subgroups based on IQ below/above 70 would have been relevant but the subgroup with IQ below 70 was too small to systematically compare it to a subgroup with IQ above 70.

Sociodemographic data As mentioned in the *methods* section assessments were funded by the Danish municipalities and assessments should therefore, in principle, be accessible to all individuals in Denmark regardless of sociodemographic background. Nevertheless, sociodemographic bias

in the municipalities' referral patterns cannot be ruled out. Therefore, the lack of sociodemographic data, beyond the sex assigned at birth, limits the transparency of the diversity in the sample. Furthermore, the lack of sociodemographic data limits the transparency of the potential influence of these factors on the diagnostic process and decision.

Conclusions

The two domains of the ADOS-2 rev. alg. (SA and RRB) were reproduced with an EFA item-to factor allocation largely replicating the one used by Hus and Lord (2014). This contributes to validating the proposed factor structure of the items in the ADOS-2 rev. alg.

For the current dataset, RFE analysis identified the item *Quality of social responses* as the second most important item for clinical diagnosis. This item was among the five most important items in a similar analysis by Küpper et al. (2020) and among the two most important items in predicting ASD diagnosis for modules 2 and 3 in an analysis by Wolff et al. (2022). Future revisions of the ADOS-2 rev. alg. could consider singling out items identified this way as mandatory for the computation of the algorithm score or perhaps using weights to adjust the importance of select items.

ROC analyses of the ADOS-2 rev. alg. for the full sample of 331 individuals resulted in an AUC of 77%, a specificity of 0.69, and a sensitivity of 0.79 (with a cut-off at the official SA + RRB score of 8). When adjusting our sample to include only individuals with an FSIQ ≥ 70 and an age ≥ 18 the AUC rose to 93%, the specificity to 0.85, and the sensitivity dropped to 0.72 (with cut-off 8). This does indicate that for an adult with FSIQ above 70, the ADOS-2 rev. alg. is a very useful instrument.

In recent years concern has been raised that the ADOS-2 is not sufficiently unisex (se Kaat et al. (2021). With this large sample of females with ASD, we do find some evidence to confirm these prior concerns since the ADOS-2 rev. alg. score is 2 points lower for females despite there being no sex differences on additional measures of FSIQ, ADI-R, or adaptive behavior. We speculate that females are better able to camouflage their challenges with social interaction during the brief ADOS-2 interview but that camouflaging behavior does not, to the same degree, affect the additional measures having a much broader scope. Nevertheless, when critically examining the mean difference between male and female ADOS-2 rev. alg. scores using the CIs, the observed mean difference could be as small as half a point. Whether or not the differences, in performance between sexes for the ADOS-2 rev. alg., are of a size necessitating a sex-specific scoring procedure is therefore difficult to judge from the

current data. Lastly, 64 female participants from our clinical sample contributed to the correlational analyses comparing the SA domain of the ADOS-2 rev. to the sum of the *Social interaction* and *Communication* domains of the ADI-R. To our knowledge, this is one of the largest groups of adult females with ASD, with scores from both ADOS-2 and ADI-R, ever reported. Seen in this light, the finding of a correlation for males but not for females does indicate a weakness in the ADOS-2 when applied to the adult female population.

Author Contributions Conceptualization: [Jens Christiansen] [Lennart Pedersen]. Methodology: [Jens Christiansen] [Lennart Pedersen]. Formal analysis and investigation: [Jens Christiansen]. Writing - original draft preparation: [Jens Christiansen]. Writing - review and editing: [Jens Christiansen] [Lennart Pedersen].

Funding Partial financial support was received from [Sofiefonden and Else Lachmanns fond].

Declarations

Ethical Approval This research study was conducted retrospectively from data obtained for clinical purposes collected after written informed consent was given. The clinic data was extracted from records the NGO was legally required to hold. During the extraction of data for this publication no identifiers entered the resulting dataset. The Danish committee system on health research ethics has been consulted extensively about this study and has no objections and testifies that the study needs no ethical approval and is legal under Danish law (reference number for waiver of ethical approval: [F-23040491]).

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

- Brugha, T. S., McManus, S., Bankart, J., Scott, F., Purdon, S., Smith, J., Bebbington, P., Jenkins, R., & Meltzer, H. (2011). Epidemiology of Autism Spectrum disorders in adults in the community in England. *Archives of General Psychiatry*, 68(5), 459. <https://doi.org/10.1001/archgenpsychiatry.2011.38>
- Fusar-Poli, L., Brondino, N., Politi, P., & Aguglia, E. (2022). Missed diagnoses and misdiagnoses of adults with autism spectrum disorder. *European Archives of Psychiatry and Clinical Neuroscience*, 272(2), 187–198. <https://doi.org/10.1007/s00406-020-01189-w>
- Hull, L., Mandy, W., & Petrides, K. (2017). Behavioural and cognitive sex/gender differences in autism spectrum condition and typically developing males and females. *Autism*, 21(6), 706–727. <https://doi.org/10.1177/1362361316669087>
- Hus, V., & Lord, C. (2014). The Autism Diagnostic Observation schedule, Module 4: Revised algorithm and standardized severity scores. *Journal of Autism and Developmental Disorders*, 44(8), 1996–2012. <https://doi.org/10.1007/s10803014-2080-3>
- Kaat, A. J., Shui, A. M., Ghods, S. S., Farmer, C. A., Esler, A. N., Thurm, A., Georgiades, S., Kanne, S. M., Lord, C., Kim, Y. S., & Bishop, S. L. (2021). Sex differences in scores on standardized measures of autism symptoms: A multisite integrative data analysis. *Journal of Child Psychology and Psychiatry*, 62(1), 97–106. <https://doi.org/10.1111/jcpp.13242>

- Kaiser, H. F., & Rice, J. (1974). Little Jiffy, Mark Iv. *Educational and Psychological Measurement*, 34(1), 111–117. <https://doi.org/10.1177/001316447403400115>
- Küpper, C., Stroth, S., Wolff, N., Hauck, F., Kliwew, N., Schad-Hansjosten, T., Kamp Becker, I., Poustka, L., Roessner, V., Schultebrucks, K., & Roepke, S. (2020). Identifying predictive features of autism spectrum disorders in a clinical sample of adolescents and adults using machine learning. *Scientific Reports*, 10(1), 4805. <https://doi.org/10.1038/s41598-020-61607-w>
- Lai, M. C., & Baron-Cohen, S. (2015). Identifying the lost generation of adults with autism spectrum conditions. *The Lancet Psychiatry*, 2(11), 1013–1027. [https://doi.org/10.1016/S2215-0366\(15\)00277-1](https://doi.org/10.1016/S2215-0366(15)00277-1)
- LeCouteur, A., Rutter, M., Lord, C., & Rios, P. (1989). Autism diagnostic interview: A standardized investigator-based instrument. *Journal of Autism and Developmental Disorders*, 19(3), 363–387.
- Leeflang, M. M. G. (2008). Systematic reviews of Diagnostic Test Accuracy. *Annals of Internal Medicine*, 149(12), 889. <https://doi.org/10.7326/0003-4819-149-12-200812160-00008>
- Lord, C., Rutter, M., & LeCouteur, A. (1994). Autism diagnostic interview—Revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders*, 24, 659–685.
- Lord, C., Rutter, M., DiLavore, P. C., & Risi, S. (1999). *Autism diagnostic observation schedule WPS (ADOS-WPS)*. Western Psychological Services.
- LordC., RisiS., LambrechtL., CookE. H., LeventhalB. L., DiLavoreP. C., PicklesA., & RutterM. (2000). The Autism Diagnostic Observation Schedule—Generic: A Standard Measure of Social and Communication Deficits Associated with the Spectrum of Autism. *Journal of Autism and Developmental Disorders*, 30(3), 205–223. <https://doi.org/10.1023/A:1005592401947>
- Lord, C., Rutter, M., DiLavore, P. C., Risi, S., Gotham, K., & Bishop, S. (2012). *Autism Diagnostic Observation schedule, second edition (ADOS-2) manual (part I): Modules 1–4*. Western Psychological Services.
- Mulherin, S. A., & Miller, W. C. (2002). Spectrum Bias or Spectrum Effect? Subgroup Variation in Diagnostic Test evaluation. *Annals of Internal Medicine*, 137(7), 598. <https://doi.org/10.7326/0003-4819-137-7-200210010-00011>
- Norman, G. R., & Streiner, D. L. (2014). *Biostatistics: The bare essentials* (4th ed.). People's Medical Publishing.
- O'Nions, E., Petersen, I., Buckman, J. E. J., Charlton, R., Cooper, C., Corbett, A., Happé, F., Manthorpe, J., Richards, M., Saunders, R., Zanker, C., Mandy, W., & Stott, J. (2023). Autism in England: Assessing underdiagnosis in a population-based cohort study of prospectively collected primary care data. *The Lancet Regional Health - Europe*, 29, 100626. <https://doi.org/10.1016/j.lanepe.2023.100626>
- Ransohoff, D. F., & Feinstein, A. R. (1978). Problems of Spectrum and Bias in evaluating the efficacy of diagnostic tests. *New England Journal of Medicine*, 299(17), 926–930. <https://doi.org/10.1056/NEJM197810262991705>
- Remnélius, K., & Bölte, S. (2023). Camouflaging in Autism: Age Effects and Cross.
- Sparrow, S., Balla, D., & Cicchetti, D. (1984). *Vineland adaptive behavior scales (surveyed)*. American Guidance Service.
- Sparrow, S., Cicchetti, V. D., & Balla, A. D. (2005). *Vineland adaptive behavior scales* (2nd ed.). American Guidance Service.
- Wechsler, D. (1981). *Manual for the Wechsler adult intelligence scale—revised*. The Psychological corporation.
- Wechsler, D. (1997). *Wechsler adult intelligence scale* (3rd ed.). The Psychological Corporation.
- Wigham, S., Rodgers, J., Berney, T., Le Couteur, A., Ingham, B., & Parr, J. R. (2019). Psychometric properties of questionnaires and diagnostic measures for autism spectrum disorders in adults: A systematic review. *Autism*, 23(2), 287–305. <https://doi.org/10.1177/1362361317748245>
- Wolff, N., Eberlein, M., Stroth, S., Poustka, L., Roepke, S., Kamp-Becker, I., & Roessner, V. (2022). Abilities and disabilities-applying machine learning to Disentangle the role of intelligence in diagnosing Autism Spectrum disorders. *Frontiers in Psychiatry*, 13, 826043. <https://doi.org/10.3389/fpsy.2022.826043>
- World Health Organization (2019). *International statistical classification of diseases and related health problems (11th ed.)* <https://icd.who.int/>
- World Health Organization. (1992). *The ICD-10 classification of mental and behavioral disorders: Clinical descriptions and diagnostic guidelines*. World Health Organization.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.