Journal of Autism and Developmental Disorders (2018) 48:2078–2089
https://doi.org/10.1007/s10803-018-3475-3

ORIGINAL PAPER

# A Comparative Analysis of the ADOS-G and ADOS-2 Algorithms: Preliminary Findings

**Taylor P. Dorlack[1] · Orrin B. Myers[2] · Piyadasa W. Kodituwakku[3]**

Published online: 27 January 2018


**Abstract**
The Autism Diagnostic Observation Schedule (ADOS) is a widely utilized observational assessment tool for diagnosis of autism spectrum disorders. The original ADOS was succeeded by the ADOS-G with noted improvements. More recently, the ADOS-2 was introduced to further increase its diagnostic accuracy. Studies examining the validity of the ADOS have produced mixed findings, and pooled relationship trends between the algorithm versions are yet to be analyzed. The current review seeks to compare the relative merits of the ADOS-G and ADOS-2 algorithms, Modules 1–3. Eight studies met inclusion criteria for the review, and six were selected for paired comparisons of the sensitivity and specificity of the ADOS. Results indicate several contradictory findings, underscoring the importance of further study.

**Keywords** Autism spectrum disorders · ADOS · Diagnostic accuracy · Diagnostic validity · Systematic literature review · Comparative analysis

## Introduction

The Autism Diagnostic Observation Schedule (ADOS) is a standardized diagnostic instrument designed to assess communication, social interaction, play skills, and restrictive and repetitive behaviors (RRB) (Lord et al. 2000). The ADOS, combined with the Autism Diagnostic Interview-Revised (ADI-R; Rutter et al. 2003) and best estimate clinical judgment, have become the accepted standards for identifying children with autism spectrum disorders (ASDs) (Mazefsky et al. 2013). Originally, the ADOS was constructed as a standardized method for direct observation of social behavior, communication, and repetitive behaviors in children who were suspected of having autism (Lord et al. 1989). The Pre-Linguistic ADOS, or PL-ADOS (DiLavore et al. 1995), was later created to extend the age and verbal limits of the original ADOS. This made the instrument more appropriate for assessing younger and/or preverbal children. Additional improvements were made in 2000, with the introduction of the ADOS-Generic, or ADOS-G. This version extended and altered the tasks administered in previous versions, leading to notable increases in diagnostic validity and reliability (Lord et al. 2000). In the most recent version of the ADOS, the ADOS-2 (Lord et al. 2012), further attempts have been made to improve the accuracy and effectiveness of the instrument.

The ADOS-2 employs revised diagnostic algorithms and modules of administration, an updated protocol for administration of the test, a new comparison score to examine overall level of autism of an individual, and reconstructed procedures revised to better align with diagnostic criteria specified by the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (Association 2013). Previously, the ADOS-G algorithms consisted of social (S), communication (C), and combined social–communication (S–C) domains. The S domain examined behaviors such as joint attention and shared enjoyment, the C domain assessed abilities such

**Electronic supplementary material** The online version of this article (https://doi.org/10.1007/s10803-018-3475-3) contains supplementary material, which is available to authorized users.

✉ Taylor P. Dorlack
tdorlack@gmail.com

[1] Department of Educational Psychology, University of Wisconsin-Madison, 1025 West Johnson Street, Madison, WI 53706, USA

[2] School of Medicine, University of New Mexico, 1 University of New Mexico Building 177, Albuquerque, NM 87131, USA

[3] Department of Pediatrics, Center for Development and Disability, University of New Mexico, 2300 Menaul Boulevard NE, Albuquerque, NM 87107, USA

as gestural use, eye contact, and facial expressions, and the combined S–C domain drew information from both individual domains (Lord et al. 2000). The ADOS-2 algorithms include domains examining social affect (SA), RRB, and combined social affect–restrictive and repetitive behaviors (SA–RRB). Here, the SA domain probes behaviors such as shared enjoyment and attention, gestural and facial expression use, and eye contact (encompassing all behaviors that were previously coded with the separate S and C domains of the ADOS-G), and the RRB domain assesses behaviors such as repetitive motor movements, preoccupation with parts of objects, and adherence to specific routines. The combined SA–RRB domain assesses information from both individual domains, similar to the S–C domain of the ADOS-G (Lord et al. 2012).

Changes to module administration options also accompany the alterations to diagnostic algorithm composition. The ADOS-G consisted of four different module administration options. Module 1 assessment was given to individuals who were preverbal or only using single words or short, simple phrases. Module 2 was administered to those who had some flexible phrase speech but were not verbally fluent. Module 3 was designed for children or adolescents with fluent verbal speech, and Module 4 was for more advanced adolescents or adults (Lord et al. 2000). By contrast, ADOS-2 contains seven distinct module options. The Toddler Module is administered to children between the ages of 12–30 months who do not consistently use phrase speech. Module 1 is still administered to individuals 31 months and older without consistent phrase speech, but is now broken into two separate administration categories: administration to those with no words and to those with only some words. During development of the ADOS-2, these novel groupings were found to improve the validity of the test (Lord et al. 2012). Module 2 is still given to those with some flexible phrase speech, but is now broken into two distinct groups for improved validity: individuals younger than 5 years old, and those older than or equal to 5 years of age. Module 3 is still administered to children and adolescents with fluent speech. Module 4, which is given to adolescents with advanced skills and adults, was not updated with the publication of the ADOS-2, although revisions have been proposed for that module (Hus and Lord 2014; Pugliese et al. 2015; de Bildt et al. 2016).

Several studies have assessed the diagnostic accuracy of the ADOS-G and ADOS-2. Regarding the ADOS-G, studies have found sensitivities ranging from 0.51 to 0.97 and specificities from 0.57 to 1.0 (Gotham et al. 2007, 2008; Wiggins and Robins 2008; de Bildt et al. 2009; Oosterling et al. 2010; Molloy et al. 2011; Kamp-Becker et al. 2011). However, because of such varied estimates of sensitivity and specificity, as well as concerns regarding the floor and ceiling effects of algorithms and the effect of impairment level (Joseph et al. 2002; de Bildt et al. 2004; Gotham et al. 2007), it became clear that further revision of the ADOS was desirable. With the creation of the revised diagnostic algorithms, later published in the ADOS-2, studies found sensitivities and specificities ranging from 0.61 to 0.97 and 0.47 to 1.0, respectively (Gotham et al. 2007, 2008; de Bildt et al. 2009; Oosterling et al. 2010; Molloy et al. 2011; Kamp-Becker et al. 2011; Zander et al. 2015). Even with the proposed improvements of ADOS-2, sensitivity and specificity variations remained apparent at the level of individual studies.

Numerous studies have examined the test performance of the ADOS-G and ADOS-2, with a few of them comparing the sensitivity and specificity of the two versions. These assessments have revealed wide variation in measured sensitivity and specificity (Lord et al. 2000; Gotham et al. 2007, 2008; Wiggins and Robins 2008; de Bildt et al. 2009; Oosterling et al. 2010; Molloy et al. 2011; Kamp-Becker et al. 2011; Zander et al. 2015). It remains unclear whether the proposed improvements of the ADOS-2 have indeed led to an improvement in diagnostic accuracy. To date, only one meta-analysis (Tsheringla et al. 2014) has been reported, which examined the diagnostic validity solely of Module 1 of the ADOS-G. Furthermore, no study has yet attempted to analyze the pooled relationship trends between the algorithms of ADOS-G and the revised algorithms of ADOS-2. Therefore, the current investigation sought to analyze the sensitivity and specificity of the ADOS-G and ADOS-2 algorithms, Modules 1–3. Modules 1–3 were chosen because of their inclusion across test versions; the Toddler Module was not included in the ADOS-G, and Module 4 of the ADOS-2 was not revised from the ADOS-G. Since we expected only a limited number of published papers on the ADOS-2, our goal was to report preliminary findings on the relative merits of the two versions, with a view of informing the design of future comparative studies of the ADOS.

## Methods

As this research focused on studies that examined the diagnostic accuracy of the ADOS-G and ADOS-2 for individuals with ASD, data were collected from studies that reported the sensitivity and specificity of the ADOS-G and ADOS-2 diagnostic assessment measures. To be eligible for additional analyses, candidate studies must have reported data that allowed for computation of sensitivity and specificity of the ADOS-G and ADOS-2 algorithms for at least one module. These criteria limited the number of studies available for analyses, but removed between-study variation which introduces confounding factors to algorithm comparisons (Leeflang et al. 2008). All methods employed for the study search procedure, inclusion and exclusion criteria, quality assessment, data extraction, and data analysis were informed

by the Preferred Reporting Items for Systematic Review and Meta-Analysis, or PRISMA (Moher et al. 2009), as well as by meta-analysis and data comparison reference guides (Borenstein et al. 2009; Field and Gillett 2010; Higgins and Green 2011).

The following electronic and print resources were searched to identify articles for possible inclusion in the study: PsychINFO, PubMed/MEDLINE, High-Wire Press, Google Scholar, Cochrane Database of Systematic Reviews, and Cochrane Controlled Trials Register. These electronic databases have been widely used in other systematic literature reviews and comparison studies. Key terms used during the electronic database searches were ASD, autism, ADOS, ADOS-G, ADOS-2, diagnostic accuracy, diagnostic validity, accuracy, validity, sensitivity, specificity, validation, psychometrics, psychometric property, factor analysis, and item response theory/Rasch model. The electronic search was not limited by any filters, and was additionally augmented with a manual search of reference lists from selected articles.

To qualify for inclusion, studies were screened using the following inclusion and exclusion criteria. First, selected studies had to compare the ability of the ADOS as the index measure and the DSM-IV/IV-TR/V, ICD-10, or best estimate clinical diagnosis as the reference measure. Second, the targeted population had to be administered the appropriate ADOS module depending on chronological age of participant and developmental level. Third, participants referred for assessment and evaluation must have met an identifying score to qualify as either "autism/ASD" or "non-autism/ASD" by the ADOS-G and ADOS-2 diagnostic algorithm cutoff scores. Fourth, clinicians administering the assessment must have received research reliability training for the ADOS to assure the reliability of test administration. Fifth, the study must have been reported in a peer-reviewed journal between 2000 and 2017. Sixth, the article must have been published in English. Seventh, selected articles must have analyzed the diagnostic validity, particularly the sensitivity and specificity, of the individual modules (Modules 1–3) of the ADOS-G or ADOS-2. Finally, chosen studies must have reported sufficient information to calculate the frequency of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) rates to support computation of sensitivity and specificity. Exclusion criteria included the following: reports published in books, conference presentations, master's theses and doctoral dissertations, correlation studies, systematic review studies, and experimental methods including single-subject or quasi-experimental designs, case studies, or controlled trials. After application of the foregoing inclusion and exclusion criteria, studies to be included in the current review were retained.

The methodological quality of each selected study was assessed using the Quality Assessment of Diagnostic Accuracy Studies 2, or QUADAS-2 (Whiting et al. 2011), by two independent reviewers. Inter-rater disagreement was resolved through discussion. Designed to assess the quality of primary diagnostic accuracy studies, the QUADAS-2 consists of domains covering participant selection, index test, reference standard, flow of participants through the study, and timing of index test and reference standard. The tool is completed in four phases: (1) the review question is stated, (2) review-specific guidance is developed, (3) each of the domains stated above is assessed for risk of bias and concerns regarding applicability, and (4) a judgment is made about the study's risk and concerns. Studies were judged as having low, high, or unclear risk of bias and concerns regarding applicability for all domains, and the quality of all articles included in the current study was assessed in this manner. Whiting et al. (2011) encourage reviewers to present a summary of the QUADAS-2 results, to highlight studies that may introduce bias to the meta-analysis, and to review all relevant evidence and investigate sources of potential heterogeneity.

Data of the selected studies were extracted using a data sheet developed in-house to gather the following information: study name and authors, algorithm type (ADOS-G and/or ADOS-2), modules included (Modules 1–3), total number of study-wide cases (subtracted participants for Module 4, if included), total number of cases per module, total number of cases classified as "autism/ASD" and "non-autism/ASD" per module (when available), number of TP, FP, FN, and TN rates for each module and algorithm, participant gender (when available), participant ethnicity (when available), whether or not studies were retrospective in nature, mean chronological age of participants (when available), chronological age range of participants (when available), and measures of diagnostic validity, including sensitivity and specificity values. Because each individual study separated autism, ASD, non-autism, and non-ASD participants into varied groupings, the current study collapsed autism and ASD into one group, and non-autism/ASD into another group. Therefore, only two groups were compared in the current study, and all reported data from individual studies were averaged within these defined groups.

All data extracted from each study were evaluated at the study-level, and individual participant data were not available for the current analyses. We began our analyses with a meta-analysis like summary of sensitivity and specificity estimates for each study, module, and algorithm. Pooled sensitivity and specificity were computed using fixed and random effects approaches (DerSimonian and Laird 1986) with SAS macros (Senn et al. 2011). Cochran's heterogeneity $Q$ statistic and the $I^2$ heterogeneity index, which is an estimate of the proportion of total variation across studies that is due to heterogeneity rather than chance (Higgins and Thompson 2002), were also computed. The primary analyses to assess ADOS-G and ADOS-2 algorithms were based

on within-study comparisons of sensitivity and specificity. Therefore, studies that did not report data from both the ADOS-G and ADOS-2 algorithms were not eligible for paired analyses. These analyses began with within-study differences in sensitivity and specificity associated with a change in algorithm, which were assessed for each module using Cochran–Mantel–Haenszel methods for a risk difference (Senn et al. 2011; Agresti 2013). The ADOS-2 versus ADOS-G differences were positive if ADOS-2 sensitivity or specificity were higher than ADOS-G proportions. The opposite was true if the differences were negative. These differences were reported for each module, study, and algorithm comparison along with 95% confidence intervals (CI), although the 95% CI did not fully account for repeated measures on children within studies. Two generalized linear mixed model approaches were used to approximate pooled estimates of ADOS-2 versus ADOS-G sensitivity or specificity differences. The first model for a binomial response variable included a random study effect, and used an identity link function to estimate the pooled average difference. The identity link was used instead of a logit link because a difference in sensitivity or specificity proportions was believed to be more interpretable than a difference in log-odds. In models like this, the random study effect accounts for repeated measures within studies. The second approach added a random effect for the algorithm difference to the previous model to estimate the amount of between study heterogeneity in the algorithm difference. Following mixed model convention, the random effects were assumed to be normally distributed, with study and algorithm effects being uncorrelated. A contrast was constructed to test whether the algorithm variance was equal to zero to assess heterogeneity. SAS v9.4 (Senn et al. 2011) was used to fit the models. Bivariate analyses of sensitivity and specificity were not conducted due to the small number of studies. We noted that data in Lord et al. (2000) (n = 168) were used in the original ADOS-G development, and that participant overlap occurred between the Lord et al. (2000) and the Gotham et al. (2007) (n = 1574) study samples. For this reason, Lord et al. (2000) was excluded from all analyses.

## Results

Database searches were conducted from October 2015–January 2016, June 2016–July 2016, May 2017–June 2017, and October 2017–November 2017. See Online Appendix A for a line-by-line literature search example. As shown in Fig. 1, there were 18,858 studies identified during initial searches of all electronic databases. 18,297 studies were immediately excluded after initial screening of titles and key words because they did not cover ASDs or the ADOS, or because they were Google Scholar duplications of studies found on PubMed/MEDLINE, PsychINFO, or High Wire Press. Next, 561 studies were identified by this search strategy, and 16 additional studies were manually retrieved (total = 577). After the application of inclusion and exclusion criteria, 569 studies were excluded. Eight studies met inclusion criteria for the review study. These studies all used the same clinical standard, namely clinical judgment based on DSM-IV or DSM-IV-TR criteria, for assessing the sensitivity and specificity of the utilized algorithms. Of them, six studies reported sufficient information on both the ADOS-G and ADOS-2 to compute sensitivity and specificity for both algorithms and were eligible for the current paired, comparative analysis. Wiggins and Robins (2008) and Zander et al. (2015) were excluded because they did not report on both the ADOS-G and ADOS-2 algorithms.

Risk of bias and applicability concerns were analyzed with the QUADAS-2 quality assessment tool. One study was rated as "unclear risk" regarding participant selection bias. As for failure to use the same reference standard in test administration, one study was rated as "unclear risk" for both reference standard and flow and timing of the study. Relating to concerns with applicability (i.e. variation in participant characterization), two studies were rated as an "unclear risk" and one study was rated as a "high risk." Additionally, one study was rated as an "unclear risk" for reference standard applicability due to issues with the reference standard. All other studies were rated as "low risk" for the various risk assessment categories. See Table 1 for more information about ratings and the rationale behind why specific studies received ratings of "unclear risk" or "high risk" for the different categories.

Sensitivity and specificity values were assessed for each module and algorithm type, and are shown in Tables 2, 3, and 4. Frequencies of TPs, FPs, TNs, and FNs, along with other study and participant characteristics, can be found in Online Appendices B and C.

For Module 1, pooled sensitivity was similar for the ADOS-G algorithm (Pooled Random Estimate = 0.87, 95% CI 0.84–0.90), the ADOS-2 algorithm administered to children with no words (0.90, 95% CI 0.87–0.93), and the ADOS-2 algorithm administered to children having some words (0.88, 95% CI 0.82–0.94, Table 2). Pooled specificity estimates for the Module 1 algorithms of the ADOS-G, ADOS-2 used with children having no words, and ADOS-2 used with children having some words were 0.71 (95% CI 0.60–0.81), 0.62 (95% CI 0.43–0.81), and 0.79 (95% CI 0.70–0.88), respectively (Table 2). Heterogeneity appeared especially prominent for the Molloy et al. (2011) study, which had low specificity across all algorithms, and for the Oosterling et al. (2010) study, which had particularly low sensitivity for the ADOS-2 algorithm used with children having some words. Paired comparisons found that specificity of the ADOS-2 algorithm used with children having

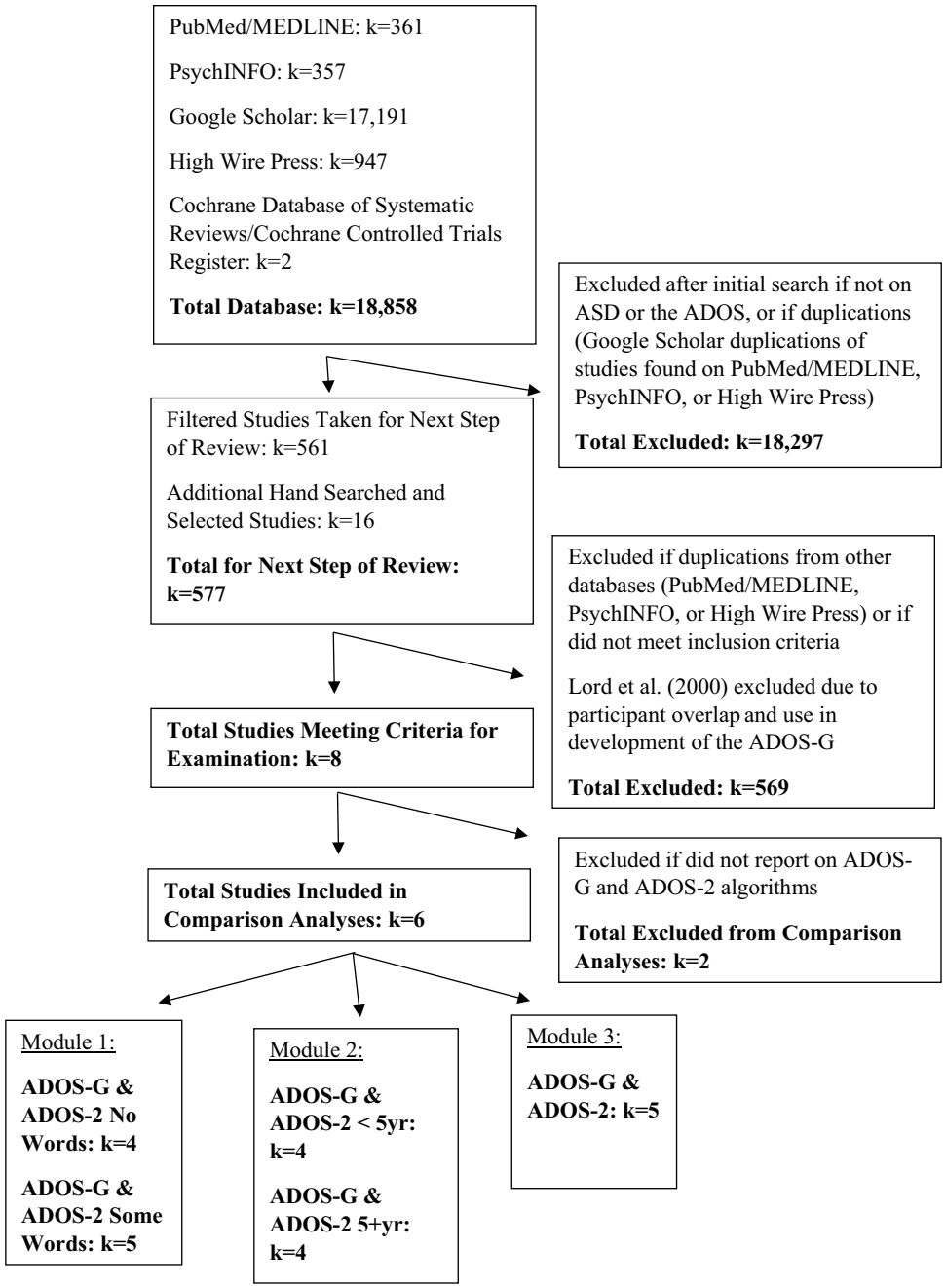


**Fig. 1** Flowchart illustrating the inclusion and exclusion of studies at each step of the systematic review

some words increased by 7% (0.07, 95% CI 0.10 to 0.25), while specificity of the ADOS-2 algorithm used with children having no words decreased by 8% (−0.08, 95% CI 0.36 to 0.21), although neither of these measures were statistically significant. Sensitivity measures remained similar across the ADOS-G and ADOS-2 algorithms, with observed changes < 3% (Table 5).

For Module 2, pooled sensitivity was 0.72 (95% CI 0.57–0.87) for the ADOS-G algorithm, 0.77 (95% CI 0.63–0.90) for the ADOS-2 algorithm administered to children < 5 years old, and 0.89 (95% CI 0.67–0.92) for the ADOS-2 algorithm administered to children older than or equal to 5 years of age (Table 3). Pooled specificity estimates for these three administrations were 0.90 (95% CI 0.83–0.97), 0.90 (95% CI 0.84–0.96), and 0.77 (95% CI 0.66–0.88), respectively (Table 3). Notable heterogeneity appeared for the Oosterling et al. (2010) study, which had low sensitivity across all algorithms, and for the Molloy et al. (2011) study, which had inconsistent sensitivity results across versions. Paired comparison analyses found that sensitivity for the ADOS-2 algorithm used with children < 5 years old remained unchanged from ADOS-G (0.01, 95% CI 0.24 to 0.26). Sensitivity of the ADOS-2 algorithm used with children 5 years and older was increased by 9%

**Table 1** Results regarding risk of bias and applicability concerns for the Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2)

| | Risk of bias | | | | Applicability concerns | | |
|---|---|---|---|---|---|---|---|
| | Participant selection | Index test | Reference standard | Flow and timing | Participant selection | Index test | Reference standard |
| Gotham et al. (2007) | Low risk | Low risk | Low risk | Low risk | Low risk | Low risk | Low risk |
| Gotham et al. (2008) | Low risk | Low risk | Low risk | Low risk | Low risk | Low risk | Low risk |
| de Bildt et al. (2009) | Unclear risk (consecutive sample of participants were not enrolled for the study—not all participants were referred for ASD) | Low risk | Unclear risk (some participants completed a diagnostic evaluation based on DSM-IV-TR criteria, while others completed a standardized classification procedure during an epidemiological study for reference standard) | Unclear risk (participants did not all receive the same reference standard) | High risk (not all participants were referred for ASD, not all received a diagnostic evaluation, and participant profiles varied significantly) | Low risk | Unclear risk (some participants completed a diagnostic evaluation based on DSM-IV-TR criteria, while others completed a standardized classification procedure during an epidemiological study) |
| Oosterling et al. (2010) | Low risk | Low risk | Low risk | Low risk | Unclear risk (participants were considered younger and higher functioning than typical samples, and the individuals evaluated more frequently presented as non-spectrum cases than ASD cases) | Low risk | Low risk |
| Molloy et al. (2011) | Low risk | Low risk | Low risk | Low risk | Low risk | Low risk | Low risk |
| Kamp-Becker et al. (2011) | Low risk | Low risk | Low risk | Low risk | Unclear risk (recruitment focused on selecting participants with high IQ scores for study participation) | Low risk | Low risk |

**Table 2** ADOS Module 1 sensitivity and specificity by algorithm (k = 5)

| Study | ADOS-G | | | | | ADOS-2 no words | | | | | ADOS-2 some words | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N | Sensitivity | 95% CI | Specificity | 95% CI | N | Sensitivity | 95% CI | Specificity | 95% CI | N | Sensitivity | 95% CI | Specificity | 95% CI |
| Gotham et al. (2007) | 847 | 0.89 | 0.87–0.91 | 0.61 | 0.52–0.69 | 495 | 0.92 | 0.90–0.95 | 0.49 | 0.35–0.63 | 352 | 0.87 | 0.83–0.91 | 0.87 | 0.79–0.94 |
| Gotham et al. (2008) | 488 | 0.87 | 0.84–0.91 | 0.84 | 0.77–0.91 | 249 | 0.86 | 0.81–0.91 | 0.80 | 0.69–0.92 | 239 | 0.92 | 0.88–0.96 | 0.83 | 0.74–0.92 |
| de Bildt et al. (2009) | 99 | 0.84 | 0.76–0.92 | 0.75 | 0.58–0.92 | | | | | | 99 | 0.89 | 0.82–0.96 | 0.67 | 0.48–0.86 |
| Oosterling et al. (2010) | 257 | 0.80 | 0.74–0.86 | 0.74 | 0.65–0.83 | 79 | 0.93 | 0.87–0.99 | 0.70 | 0.42–0.98 | 178 | 0.72 | 0.64–0.81 | 0.86 | 0.79–0.94 |
| Molloy et al. (2011) | 177 | 0.89 | 0.84–0.94 | 0.57 | 0.43–0.71 | 87 | 0.87 | 0.79–0.95 | 0.47 | 0.23–0.71 | 90 | 0.97 | 0.92–1.00 | 0.56 | 0.39–0.73 |
| Pooled fixed | | 0.88 | 0.86–0.89 | 0.73 | 0.69–0.77 | | 0.91 | 0.89–0.93 | 0.66 | 0.58–0.74 | | 0.90 | 0.88–0.92 | 0.83 | 0.78–0.87 |
| Pooled random | | 0.87 | 0.84–0.90 | 0.71 | 0.60–0.81 | | 0.90 | 0.87–0.93 | 0.62 | 0.43–0.81 | | 0.88 | 0.82–0.94 | 0.79 | 0.70–0.88 |
| $I^2$ | | 57.5 | | 81.9*** | | | 49.6 | | 79.0** | | | 84.5*** | | 71.2** | |

Heterogeneity test results: **p < 0.01; ***p < 0.001

**Table 3** ADOS Module 2 sensitivity and specificity by algorithm (k = 5)

| Study | ADOS-G | | | | | ADOS-2 < 5 years old | | | | | ADOS-2 ≥ 5 years old | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N | Sensitivity | 95% CI | Specificity | 95% CI | N | Sensitivity | 95% CI | Specificity | 95% CI | N | Sensitivity | 95% CI | Specificity | 95% CI |
| Gotham et al. (2007) | 329 | 0.89 | 0.85–0.93 | 0.85 | 0.76–0.94 | 137 | 0.91 | 0.85–0.96 | 0.87 | 0.75–0.99 | 192 | 0.91 | 0.86–0.95 | 0.87 | 0.75–0.99 |
| Gotham et al. (2008) | 87 | 0.87 | 0.79–0.95 | 1.00 | 0.81–1.00 | 87 | 0.80 | 0.70–0.89 | 1.00 | 0.81–1.00 | | | | | |
| de Bildt et al. (2009) | 124 | 0.59 | 0.48–0.71 | 0.78 | 0.67–0.89 | | | | | | 124 | 0.70 | 0.60–0.81 | 0.70 | 0.57–0.83 |
| Oosterling et al. (2010) | 203 | 0.51 | 0.43–0.60 | 0.96 | 0.91–1.00 | 105 | 0.61 | 0.50–0.72 | 0.83 | 0.70–0.97 | 98 | 0.64 | 0.51–0.76 | 0.85 | 0.74–0.96 |
| Molloy et al. (2011) | 198 | 0.73 | 0.65–0.82 | 0.84 | 0.76–0.91 | 107 | 0.72 | 0.60–0.84 | 0.70 | 0.57–0.82 | 91 | 0.90 | 0.82–0.98 | 0.63 | 0.47–0.78 |
| Pooled fixed | | 0.81 | 0.78–0.83 | 0.90 | 0.86–0.93 | | 0.82 | 0.78–0.86 | 0.80 | 0.73–0.87 | | 0.86 | 0.83–0.90 | 0.78 | 0.72–0.84 |
| Pooled random | | 0.72 | 0.57–0.87 | 0.90 | 0.83–0.97 | | 0.77 | 0.63–0.90 | 0.90 | 0.84–0.96 | | 0.89 | 0.67–0.92 | 0.77 | 0.66–0.88 |
| $I^2$ | | 95.2*** | | 70.4** | | | 88.3*** | | 25.1 | | | 88.7*** | | 67.1* | |

Heterogeneity test results: *p < 0.05; **p < 0.01; ***p < 0.001

**Table 4** ADOS Module 3 sensitivity and specificity by algorithm (k = 5)

| Study | ADOS-G | | | | | ADOS-2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | N | Sensitivity | 95% CI | Specificity | 95% CI | N | Sensitivity | 95% CI | Specificity | 95% CI |
| Gotham et al. (2007) | 398 | 0.77 | 0.73–0.82 | 0.83 | 0.75–0.91 | 398 | 0.82 | 0.77–0.86 | 0.80 | 0.71–0.88 |
| Gotham et al. (2008) | 457 | 0.60 | 0.56–0.65 | 0.93 | 0.87–0.99 | 457 | 0.71 | 0.67–0.76 | 0.90 | 0.84–0.97 |
| de Bildt et al. (2009) | 335 | 0.69 | 0.63–0.74 | 0.76 | 0.66–0.86 | 335 | 0.77 | 0.72–0.82 | 0.69 | 0.58–0.79 |
| Molloy et al. (2011) | 209 | 0.82 | 0.75–0.90 | 0.60 | 0.51–0.69 | 209 | 0.90 | 0.83–0.96 | 0.45 | 0.36–0.54 |
| Kamp-Becker et al. (2011) | 252 | 0.87 | 0.81–0.92 | 0.81 | 0.74–0.88 | 252 | 0.93 | 0.88–0.97 | 0.75 | 0.68–0.83 |
| Pooled fixed | | 0.73 | 0.71–0.76 | 0.82 | 0.79–0.85 | | 0.82 | 0.80–0.84 | 0.75 | 0.71–0.79 |
| Pooled random | | 0.75 | 0.66–0.84 | 0.79 | 0.68–0.90 | | 0.82 | 0.75–0.90 | 0.72 | 0.57–0.87 |
| $I^2$ | | 92.9*** | | 89.6*** | | | 92.6*** | | 93.7*** | |

Heterogeneity test results: ***$p < 0.001$

from the ADOS-G (0.09, 95% CI 0.03 to 0.21), although it was not statistically significant. Specificity measures were reduced by 8% for the ADOS-2 algorithm used with children < 5 years old (− 0.08, 95% CI 0.21 to 0.05) and by 10% for the ADOS-2 algorithm used with children 5 years of age and older (− 0.10, 95% CI 0.28 to 0.07). However, none of these differences were statistically significant (Table 5).

For Module 3, pooled sensitivity was 0.75 (95% CI 0.66–0.84) for the ADOS-G algorithm and 0.82 (95% CI 0.75–0.90) for the ADOS-2 algorithm (Table 4). Pooled specificity estimates for the Module 3 ADOS-G and ADOS-2 algorithms were 0.79 (95% CI 0.68–0.90) and 0.72 (95% CI 0.57–0.87), respectively (Table 4). Prominent heterogeneity appeared for the Molloy et al. (2011) study, which had particularly low specificity estimates for both the ADOS-G and ADOS-2 algorithms. Paired comparison analyses found that sensitivity of the ADOS-2 algorithm was significantly improved by 8% (0.08, 95% CI 0.03–0.13) from ADOS-G. Specificity of the ADOS-2 algorithm was decreased by 7% (− 0.07, 95% CI 0.26 to 0.12) from ADOS-G, although it was not significant (Table 5).

## Discussion

The main objective of the current investigation was to evaluate the relative merits of the ADOS-G and ADOS-2 algorithms for Modules 1–3. The sensitivity and specificity of the ADOS-G and ADOS-2 were examined directly in paired comparison analyses of studies reporting on both versions. However, the number of studies matching our inclusion criterion was small, with only four to five studies available for analyses of any module.

For Module 1, pooled sensitivity remained essentially unchanged across the ADOS-G and ADOS-2 algorithms. Pooled specificity measures were more variable, with estimates for the ADOS-2 algorithm used with children having no words decreasing, and those for the ADOS-2 algorithm used with children having some words increasing, from the ADOS-G algorithm. Paired comparison analyses found that specificity of the ADOS-2 algorithm tended to increase from that of the ADOS-G when used with children having some words and decrease when used with children having no words, although both findings were insignificant. Sensitivity remained essentially the same across versions.

When examining algorithm changes between the ADOS-G and ADOS-2, there are several possible explanations for why pooled sensitivity may have remained constant between the ADOS-G and ADOS-2 algorithms, and why our paired comparison analyses showed consistent sensitivity between the two ADOS versions. First, the separate social (S) and communication (C) sections previously assessed with the ADOS-G algorithms are now combined into the SA section of the ADOS-2. Therefore, the same social communication items are measured with the ADOS-2 that were already accounted for by the ADOS-G. Additionally, some correlational results from the validity reports for the ADOS-2 (Lord et al. 2012) may account for the lack of change in sensitivity for Module 1, as there were reports of item correlation inconsistencies regarding the SA and restricted and repetitive behaviors (RRB) domains of the algorithms. For the Module 1 algorithm used with children having some words, the "stereotyped/idiosyncratic use of words or phrases" item was shown to correlate equally with both the SA and RRB domains. Additionally, for the Module 1 assessment administered to children with no words, the "intonation" item correlated highly with nonverbal mental age in all age categorizations, indicating that language development is a correlational factor. Second, non-verbal and verbal mental age correlations were reported for items in Module 1. When utilized with children having nonverbal mental ages > 15 months, the "intonation" and "repetitive interests" items of the Module 1 assessment for children having no words were shown to correlate more with the SA domain than the intended RRB domain. These demonstrate that the ADOS remains language-dependent, and indicate a strong

**Table 5** Within study differences between ADOS-2 and ADOS-G (referent) for sensitivity and specificity of autism diagnosis by module

| Module | Study | [ADOS-2 no words—ADOS-G] | | [ADOS-2 some words—ADOS-G] | |
|---|---|---|---|---|---|
| | | Sensitivity | Specificity | Sensitivity | Specificity |
| | | Difference (95% CI) | Difference (95% CI) | Difference (95% CI) | Difference (95% CI) |
| 1 | Gotham et al. (2007) | 0.03 (−0.00 to 0.06) | −0.12 (−0.28 to 0.05) | −0.02 (−0.07 to 0.02) | 0.26 (0.15 to 0.37) |
| | Gotham et al. (2008) | −0.01 (−0.07 to 0.05) | −0.03 (−0.17 to 0.10) | 0.05 (−0.01 to 0.10) | −0.01 (−0.12 to 0.11) |
| | de Bildt et al. (2009) | | | 0.05 (−0.06 to 0.16) | −0.08 (−0.34 to 0.17) |
| | Oosterling et al. (2010) | 0.13 (0.04 to 0.22) | −0.04 (−0.34 to 0.25) | −0.07 (−0.18 to 0.04) | 0.12 (0.00 to 0.24) |
| | Molloy et al. (2011) | −0.02 (−0.11 to 0.08) | −0.10 (−0.38 to 0.17) | 0.07 (0.00 to 0.15) | −0.01 (−0.23 to 0.21) |
| | Pooled fixed | 0.03 (−0.02 to 0.07) | −0.08 (−0.20 to 0.04) | 0.00 (−0.04 to 0.05) | 0.10 (0.00 to 0.19) |
| | Pooled random | 0.03 (−0.03 to 0.08) | −0.08 (−0.36 to 0.21) | 0.01 (−0.06 to 0.08) | 0.07 (−0.10 to 0.25) |
| | Heterogeneity p-value | 0.87 | 0.13 | 0.14 | 0.02 |
| | # Studies | k = 4 | k = 4 | k = 5 | k = 5 |
| | | [ADOS-2 < 5year—ADOS-G] | | [ADOS-2 5+ year—ADOS-G] | |
| 2 | Gotham et al. (2007) | 0.02 (−0.05 to 0.08) | 0.02 (−0.13 to 0.17) | 0.02 (−0.04 to 0.08) | 0.02 (−0.13 to 0.17) |
| | Gotham et al. (2008) | −0.07 (−0.20 to 0.05) | 0.00 (−0.10 to 0.10) | | |
| | de Bildt et al. (2009) | | | 0.11 (−0.04 to 0.26) | −0.08 (−0.25 to 0.09) |
| | Oosterling et al. (2010) | 0.10 (−0.04 to 0.24) | −0.12 (−0.27 to 0.02) | 0.13 (−0.02 to 0.28) | −0.11 (−0.23 to 0.01) |
| | Molloy et al. (2011) | −0.01 (−0.16 to 0.14) | −0.14 (−0.29 to 0.00) | 0.17 (0.05 to 0.29) | −0.21 (−0.38 to −0.05) |
| | Pooled fixed | 0.02 (−0.08 to 0.12) | −0.08 (−0.20 to 0.04) | 0.08 (−0.02 to 0.17) | −0.11 (−0.33 to 0.05) |
| | Pooled random | 0.01 (−0.24 to 0.26) | −0.08 (−0.21 to 0.05) | 0.09 (−0.03 to 0.21) | −0.10 (−0.28 to 0.07) |
| | Heterogeneity p-value | 0.005 | 0.43 | 0.24 | 0.38 |
| | # Studies | k = 4 | k = 4 | k = 4 | k = 4 |
| | | [ADOS-2—ADOS-G] | | | |
| 3 | Gotham et al. (2007) | 0.04 (−0.02 to 0.11) | −0.04 (−0.15 to 0.08) | | |
| | Gotham et al. (2008) | 0.11 (0.04 to 0.17) | −0.03 (−0.12 to 0.06) | | |
| | de Bildt et al. (2009) | 0.09 (0.01 to 0.16) | −0.07 (−0.22 to 0.08) | | |
| | Molloy et al. (2011) | 0.07 (−0.02 to 0.17) | −0.15 (−0.28 to −0.02) | | |
| | Kamp-Becker et al. (2011) | 0.06 (−0.01 to 0.14) | −0.06 (−0.16 to to 0.05) | | |
| | Pooled fixed | **0.08 (0.03 to 0.13)** | −0.07 (−0.15 to 0.01) | | |
| | Pooled random | **0.08 (0.03 to 0.13)** | −0.07 (−0.26 to 0.12) | | |
| | Heterogeneity p-value | 0.86 | 0.08 | | |
| | # Studies | k = 5 | k = 5 | | |

Mantel–Haenszel differences for proportions are shown for individual studies. Pooled estimates account for within-study repeated measures, and pooled random effect estimates account for between-study variation and within-study repeated measures. Heterogeneity was assessed by testing whether the variance of between-study algorithm differences was greater than zero for the pooled random effects model. Pooled algorithm differences that are significantly different from zero are in boldface

likelihood that social and language abilities are interrelated. The authors also reported that six different items correlated highly with verbal mental age for the Module 1 assessment administered to children having some words. Further, the ADOS diagnostic classifications of "autism," "non-autism ASD," and "non-spectrum" have remained consistent between both versions of the assessment, and would likely classify children the same way. Pooled specificity appeared to increase with administration of the ADOS-2 algorithm used with children having some words and decrease with the ADOS-2 algorithm used with children having no words from the ADOS-G, but these differences were found to be insignificant during paired comparison analyses. Thus, although trends can be identified during examination of the pooled data, the measure's ability to accurately detect when children do not qualify as having autism seems to have remained the same across ADOS versions.

For Module 2, pooled sensitivity appeared to increase from the ADOS-G algorithm with both ADOS-2 administrations. Pooled specificity remained constant between the ADOS-G and ADOS-2 algorithm used with children younger than 5 years old, and decreased for the ADOS-2 algorithm used with children age 5 years and older from ADOS-G. Paired comparison analyses found that sensitivity

remained constant between the ADOS-G and ADOS-2 algorithm administered to children under 5 years old, and sensitivity of the ADOS-2 algorithm administered to children age 5 years and older suggested an increase from the ADOS-G algorithm. Specificity estimates of both ADOS-2 algorithms suggested a decrease from the ADOS-G algorithm. However, all paired comparison trends were insignificant, indicating that diagnostic accuracy has remained essentially constant across ADOS versions.

In validity reports for ADOS-2, Lord et al. (2012) mentioned correlational results that may lend support to these findings. First, an item correlation inconsistency for the Module 2 algorithm used with children greater than or equal to 5 years of age reflects a potential inability of the current SA and RRB domains to distinguish measured behaviors. They report that the "stereotyped/idiosyncratic use of words or phrases" item correlated higher with the SA domain than the intended RRB domain. As with the Module 1 algorithms, the revised Module 2 algorithms may be as language dependent as those of the ADOS-G. For the Module 2 assessment administered to children younger than 5 years of age, the "gestures" and "unusual eye contact" items were highly correlated with chronological age, indicating the module's dependence on developed communication abilities. Again, the S and C domains previously associated with the ADOS-G are now collapsed into one SA domain. The same key items are being measured with ADOS-2 as in ADOS-G, and the same diagnostic classifications are being used to separate "autism," "non-autism ASD," and "non-spectrum." In the absence of alterations to assessment items within modules and new diagnostic classifications that align with current DSM-V diagnostic criteria, children are likely to be assessed and classified in the same ways, with sensitivity and specificity remaining essentially constant. These inconsistencies and correlational results may provide rationale for the measure's unchanged ability to determine children who do and do not have ASD.

For Module 3, pooled sensitivity increased, and pooled specificity decreased, for the ADOS-2 algorithm from that of the ADOS-G algorithm. Paired comparison analyses found a statistically significant increase in sensitivity from ADOS-G with the ADOS-2 algorithm. Specificity estimates of the ADOS-2 algorithm indicated a decrease from ADOS-G, although this finding was insignificant.

The significant increase in sensitivity with administration of the ADOS-2 is important to note when examining results for Module 3. Because fluent speech is used as a qualifier for the administration of Module 3, there is less chance for correlation of items on the revised version with language abilities. This allows for increased sensitivity in this module. Changes in algorithm composition within the SA and RRB domains seem to have enhanced the ability to distinguish individuals qualifying for diagnosis. Results also indicate essentially unchanged specificity between the ADOS-2 and ADOS-G algorithms. Here, although the new SA and RRB domains of the ADOS-2 seem to have increased accurate determination of individuals who qualify for a diagnosis of autism, the measure still employs the same diagnostic categories for autism as the ADOS-G and appears to capture children who do not qualify for a diagnosis of autism in similar ways.

In addition to the rationale behind our comparative findings, authors who have examined the ADOS-G and ADOS-2 have also cited numerous reasons for variability in diagnostic performance. This includes varied sample composition and functioning levels between study participants (Gotham et al. 2008; de Bildt et al. 2009; Oosterling et al. 2010; Molloy et al. 2011; Kamp-Becker et al. 2011; Zander et al. 2015), overlap in scores for ADOS items between children with autism/ASD and other neurodevelopmental and psychiatric disorders (Bishop et al. 2007; Gotham et al. 2007; Klein-Tasman et al. 2007; Leyfer et al. 2008; Molloy et al. 2011), variation in test administration and reliability of the examiner (Lord et al. 2000; Gotham et al. 2007, 2008; de Bildt et al. 2009; Oosterling et al. 2010; Kamp-Becker et al. 2011), errors in coding of the ADOS (Gotham et al. 2008; Oosterling et al. 2010), and incorrect choice of administration module (Gotham et al. 2007; Oosterling et al. 2010). These factors may have contributed to our results and may account for why we see such relatively high levels of heterogeneity throughout the data. Furthermore, the assessment of test performance involves use of clinical diagnosis as the gold standard, which is known to vary depending on the diagnostic criteria employed and the clinician's competence and experience.

## Limitations

Several limitations with the current review should be noted. First, the total number of studies included in the review was rather small due to limited research having been conducted in the area of interest. Only eight studies met criteria for inclusion (with the exclusion of Lord et al. 2000), and even smaller numbers provided information used to calculate pooled sensitivity values, pooled specificity values, and paired comparisons for each of the modules examined. A total of five studies each provided pooled sensitivity and specificity information for Modules 1, 2, and 3. Furthermore, even smaller numbers of studies were included during paired comparison analyses for the individual algorithms. Such a limited collection of studies may lead to issues with sampling error and effect size, and our results should be viewed as preliminary findings. Second, the quality of studies as assessed by the QUADAS-2 substantially varied. Although most of the included studies presented low risk of bias and few applicability concerns regarding the various

areas of consideration, some studies presented with high or unclear risks. In Module 1, two studies (de Bildt et al. 2009; Oosterling et al. 2010) presented high or unclear risks of bias and applicability concerns related to participant selection, reference standard, and flow and timing of the study. In Module 2, two studies (de Bildt et al. 2009; Oosterling et al. 2010) presented high or unclear risks of bias and applicability concerns related to participant selection, reference standard, and flow and timing of the research study. In Module 3, two studies (de Bildt et al. 2009; Kamp-Becker et al. 2011) presented high or unclear risks of bias and applicability concerns, again related to participant selection, reference standard, and flow and timing of the study. The inclusion of studies with risks related to quality may introduce additional heterogeneity and bias to the study, potentially influencing outcomes and making conclusions more difficult to draw. Third, statistical heterogeneity was relatively high across all pooled sensitivity and specificity values, implying variability in effects between studies caused by methodological differences. Fourth, evaluation of diagnostic accuracy of the ADOS was dependent on information reported in published articles. It is possible that selective reporting may have occurred within studies, and that publication bias may have affected the stated results. Fifth, the methodological choice to combine autism/ASD into one group, and non-autism/ASD into another, may have impacted the direction of results in ways that are unknown. Additionally, the small number of available studies limited our ability to conduct bivariate analyses of sensitivity and specificity to account for their correlation. As more studies become available, additional analyses may be warranted to explore sources of heterogeneity. Further, no studies to date have examined the diagnostic accuracy of individual modules of the ADOS in comparison to recently published DSM-V diagnostic criteria. Studies in the current review utilized the ADOS-G and ADOS-2 in comparison to DSM-IV or DSM-IV-TR criteria, providing a limited scope of analysis. Because clinicians are now using the ADOS-2 and DSM-V diagnostic criteria during evaluation of ASDs, this data would be of significant interest and have the largest impact on current diagnostic practices.

### Future Directions

Despite the previously mentioned limitations, the current analysis highlights the importance of future research that examines and seeks to improve the diagnostic validity of the ADOS. While results of the current study show that sensitivity and specificity appear to have increased in some cases from ADOS-G with administration of the ADOS-2, other instances show equivalent or marginally decreased diagnostic accuracy with the latest version of the assessment. These results indicate that the effectiveness of module administration and algorithm revisions in improving the validity of the measure has been unclear, and that further research and assessment of the ADOS is needed. Potential revisions to consider include ways to make the assessment less dependent on language capabilities and better able to examine nonverbal communication, especially in Modules 1 and 2, as well as better aligning the measure's diagnostic categorizations with the current DSM-V diagnostic criteria for ASD. Additionally, future research must examine the diagnostic accuracy of the ADOS with DSM-V diagnostic criteria as the reference standard to best inform our knowledge and diagnosis of ASDs. Only when such changes are made can the diagnostic accuracy of the instrument be fully assessed and potential points of improvement noted.

**Acknowledgments** This work was partially supported by the National Center for Advancing Translational Sciences of the National Institutes of Health (UL1TR001449).

**Author Contributions** TPD conceived of the study, participated in its design and coordination, performed the measurement, participated in statistical analysis and interpretation of data, and drafted/edited the manuscript. OBM performed the statistical analysis, interpreted the data, and helped to draft/edit the manuscript. PWK participated in its design and coordination, and helped to draft/edit the manuscript. All authors read and approved the final manuscript.

### Compliance with Ethical Standards

**Conflict of interest** The authors declare that they have no conflict of interest.

**Research Involving Human and Animal Rights** This article does not contain any studies with human participants or animals performed by any of the authors.

## References

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Hoboken, NJ: Wiley.

Association, A. P. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

Bishop, S., Gahagan, S., & Lord, C. (2007). Re-examining the core features of autism: A comparison of autism spectrum disorder and fetal alcohol spectrum disorder. *The Journal of Child Psychology and Psychiatry, 48*(11), 1111–1121.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., et al. (2009). *Introduction to meta-analysis*. Chichester: Wiley.

de Bildt, A., Sytema, S., Ketelaars, C., et al. (2004). Interrelationship between Autism Diagnostic Observation Schedule-Generic (ADOS-G), Autism Diagnostic Interview-Revised (ADI-R), and the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV-TR) classification in children and adolescents with mental retardation. *Journal of Autism and Developmental Disorders, 34*(2), 129–137.

de Bildt, A., Sytema, S., Meffert, H., et al. (2016). The Autism Diagnostic Observation Schedule, module 4: Application of the revised algorithms in an independent, well-defined, Dutch sample

(n = 93). *Journal of Autism and Developmental Disorders, 46*(1), 21–30.

de Bildt, A., Sytema, S., van Lang, N. D., et al. (2009). Evaluation of the ADOS revised algorithm: The applicability in 558 Dutch children and adolescents. *Journal of Autism and Developmental Disorders, 39*(9), 1350–1358.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials, 7*, 177–188.

DiLavore, P. C., Lord, C., & Rutter, M. (1995). The pre-linguistic autism diagnostic observation schedule. *Journal of Autism and Developmental Disorders, 25*(4), 355–379.

Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology, 63*(3), 665–694.

Gotham, K., Risi, S., Dawson, G., et al. (2008). A replication of the Autism Diagnostic Observation Schedule (ADOS) revised algorithms. *Journal of the American Academy of Child and Adolescent Psychiatry, 47*(6), 642–651.

Gotham, K., Risi, S., Pickles, A., et al. (2007). The Autism Diagnostic Observation Schedule: Revised algorithms for improved diagnostic validity. *Journal of Autism and Developmental Disorders, 37*(4), 613–627.

Higgins, J. P. T., & Green, S. (2011). *Cochrane handbook for systematic reviews of interventions (version 5.1.0)*. The Cochrane Collaboration.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*, 1539–1558.

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.

Joseph, R. M., Tager-Flusberg, H., & Lord, C. (2002). Cognitive profiles and social-communicative functioning in children with autism spectrum disorder. *Journal of Child Psychology and Psychiatry, 43*(6), 807–821.

Kamp-Becker, I., Ghahreman, M., Heinzel-Gutenbrunner, M., et al. (2011). Evaluation of the revised algorithm of Autism Diagnostic Observation Schedule (ADOS) in the diagnostic investigation of high-functioning children and adolescents with autism spectrum disorders. *Autism, 17*(1), 87–102.

Klein-Tasman, B. P., Mervis, C. B., Lord, C., et al. (2007). Socio-communicative deficits in young children with Williams syndrome: Performance on the Autism Diagnostic Observation Schedule. *Child Neuropsychology, 13*(5), 444–467.

Leeflang, M. M. G., Deeks, J. J., Gatsonis, C., et al. (2008). Systematic reviews of diagnostic test accuracy. *Annals of Internal Medicine, 149*(12), 889–897.

Leyfer, O. T., Tager-Flusberg, H., Dowd, M., et al. (2008). Overlap between autism and specific language impairment: Comparison of Autism Diagnostic Interview and Autism Diagnostic Observation Schedule scores. *Autism Research, 1*(5), 284–296.

Lord, C., Risi, S., Lambrecht, L., et al. (2000). The Autism Diagnostic Observation Schedule-Generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders, 30*(3), 205–223.

Lord, C., Rutter, M., DiLavore, P. C., et al. (2012). *Autism diagnostic observation schedule* (3rd ed.). Torrance, CA: Western Psychological Services.

Lord, C., Rutter, M., Goode, S., et al. (1989). Autism diagnostic observation schedule: A standardized observation of communicative and social behavior. *Journal of Autism and Developmental Disorders, 19*(2), 185–212.

Mazefsky, C. A., McPartland, J. C., Gastgeb, H. Z., et al. (2013). Brief report: Comparability of DSM-IV and DSM-V ASD research samples. *Journal of Autism and Developmental Disorders, 43*(2), 1236–1242.

Moher, D., Liberati, A., Tetzlaff, J., et al. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine, 151*(4).

Molloy, C. A., Murray, D. S., Akers, R., et al. (2011). Use of the Autism Diagnostic Observation Schedule (ADOS) in a clinical setting. *Autism, 15*(2), 143–162.

Oosterling, I., Roos, S., de Bildt, A., et al. (2010). Improved diagnostic validity of the ADOS revised algorithms: A replication study in an independent sample. *Journal of Autism and Developmental Disorders, 40*(6), 689–703.

Pugliese, C. E., Kenworthy, L., Bal, V. H., et al. (2015). Replication and comparison of the newly proposed ADOS-2, module 4 algorithm in ASD without ID: A multi-site study. *Journal of Autism and Developmental Disorders, 45*(12), 3919–3931.

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

Senn, S., Weir, J., Hua, T. A., et al. (2011). Creating a suite of macros for meta-analysis in SAS: A case study in collaboration. *Statistics & Probability Letters, 81*(7), 842–851.

Tsheringla, S., Minju, K. A., Russell, S., et al. (2014). A meta-analysis of the diagnostic accuracy of Autism Diagnostic Observation Schedule module-1 for autism spectrum disorders. *Indian Journal of Pediatrics, 81*(Suppl 2), 187–192.

Whiting, P. F., Rutjes, A. W., Westwood, M. E., et al. (2011). QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine, 155*(8), 529–536.

Wiggins, L. D., & Robins, D. L. (2008). Brief report: Excluding the ADI-R behavioral domain improves diagnostic agreement in toddlers. *Journal of Autism and Developmental Disorders, 38*(5), 972–976.

Zander, E., Sturm, H., & Bolte, 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.