



Diagnostic Utility of the Gilliam Autism Rating Scales-3rd Edition Parent Report in Clinically Referred Children

Amy Camodeca^{1,2,3}

Accepted: 9 February 2022 / Published online: 4 March 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

There is limited research regarding the Gilliam Autism Rating Scales-3rd Edition (GARS-3) despite its extensive use. A comprehensive diagnostic evaluation, including the Autism Diagnostic Observation Schedule-2nd Edition (ADOS-2) was provided to 186 clinically referred children suspected of autism (\bar{X} age = 8.98; Autism [AUT] $n = 87$; Not Autism [NOT] $n = 99$). Mean difference analyses, Logistic Regressions, and ROC analyses were non-significant for both Autism Index scores. The author-suggested cutoff score of 70 correctly classified approximately 47% of participants, with false positive rates = 82.83–87.88%. ADOS-2 correlations were significantly lower vis-à-vis the standardization sample. The Social Interaction subscale demonstrated weak, marginal results, and sensitivity/specificity could not be optimized. In its current form, the GARS-3 does not demonstrate adequate criterion validity for use in assessment of complex community samples.

Keywords Gilliam Autism Rating Scales · Autism spectrum disorder · Validity · Assessment · Diagnosis

Introduction

Children and adults with autism spectrum disorder (ASD) have deficits in social communication and cognitive/behavioral flexibility (American Psychiatric Association, 2013). The incidence of ASD has increased in the last two decades, as has the desire for early and accurate diagnosis (Basiru et al., 2021; Davidovitch et al., 2020; Hamad et al., 2019). Multiple ASD symptom questionnaires have been published to assist in screening and diagnosis of ASD, but little research exists on the psychometric properties of these measures (Charman & Gotham, 2013; Hampton & Strand, 2015; Wigham et al., 2019).

One such measure is the Gilliam Autism Rating Scale-3rd Edition (GARS-3; Gilliam, 2014). The GARS-3 is widely utilized in clinical practice. Over 550 GARS-3 kits and more than 2000 GARS-3 form packages were purchased in 2020

(Cooter, 2021, personal communication). Additionally, researchers have utilized the GARS-3 in diagnostic classification, (Alsaedi et al., 2020; Lordo et al., 2017; Pfeiffer et al., 2018), educational classification (Cardon et al., 2019) and quantitative ASD symptom measurement (Knowland et al., 2019; Tse et al., 2018). However, aside from the standardization sample, there is no research attesting to the utility of the GARS-3 in ASD diagnosis or ability to measure ASD traits.

The GARS-3 Measure and Standardization Sample

The GARS-3 (Gilliam, 2014) is a 58-item Likert scale questionnaire designed for either parent or teacher report of persons aged 3–22 suspected of ASD. According to the author, the GARS-3 improves on the GARS-2 (Gilliam, 2006) regarding content validity and congruence with DSM-5 diagnostic criteria by including newly-developed questions and questions identified from other ASD measures. There are six subscales (Restrictive/Repetitive Behavior, Social Interaction, Social Communication, Emotional Responses, Cognitive Style, and Maladaptive Speech) and two Autism Index (AI) scores. One AI is comprised of the first four subscales listed above and one AI is comprised of all six subscales. The four- and six-subscale AI scores are referred to in this paper as the AI-4 and AI-6, respectively. The AI-4 is

✉ Amy Camodeca
asc19@psu.edu

¹ Psychology Department, The Pennsylvania State University,
100 University Drive, Monaca, PA 15061, USA

² University of Windsor, Windsor, ON, Canada

³ Present Address: The Pennsylvania State University, Beaver
Campus, 100 University Drive, Monaca, PA 15108, USA

appropriate for children who are non-verbal, as the Cognitive Style and Maladaptive Speech questions would not apply to them. The AI-6 is appropriate for verbal children. To score the GARS-3, responses on each subscale are summed to create a raw score, which is then converted to a scaled score ($\bar{X} = 10$, $SD = 3$). The first four or all six scaled scores are summed and converted to AI-4 and AI-6 standard scores, respectively ($\bar{X} = 100$, $SD = 15$) (Gilliam, 2014).

The GARS-3 provides cutoff points regarding likelihood of ASD diagnosis and Level of Severity based on the AI-4/AI-6 score. Scores ≤ 54 are considered Unlikely/Not ASD, and no supports needed; 55–70 is Probable, Level 1 ASD, and requiring minimal supports; 71–100 is Very Likely, Level 2 ASD, and requiring substantial supports; ≥ 101 is also considered Very Likely, but classified as Level 3 ASD and requiring very substantial supports (Gilliam, 2014). The normative scores (and cutoffs) were created based on a standardization sample of persons with ASD. Thus a score in the impaired range does not reflect impairment in that skill, but the proportion of the normative sample scoring in those areas. For example, 0.2% of the normative sample scored below 55, and 4% scored between 55 and 70 (Gilliam, 2014).

The standardization sample included 1,859 children and young adults with ASD from across the United States. The majority (61.3%) had ASD only, and the remaining people had ASD and unspecified comorbid diagnoses. The sample was comprised of mostly (77%) males, which the author specified was consistent with the male:female ratio in ASD. The sample was predominately white (80%), consistent with the proportions of the school-age population at the time. The scores are not normed for age or gender, although only age was investigated as a potential confounding variable by the author. Correlations with age were very weak for all scaled scores, $r_s \leq .16$; correlations with Index scores were not reported (Gilliam, 2014).

Criterion Validity

Criterion validity was established via correlations with two autism questionnaires, an observational autism rating scale, and one gold-standard measure, the Autism Diagnostic Observation Schedule (ADOS). For analyses, the ADOS Communication + Social raw scores were transformed to standard scores for correlation with the GARS-3. The author did not provide a rationale for this conversion, but it is likely that this was done in order to include participants from all ADOS modules in analyses, as algorithms for different module produce raw scores of different ranges. Correlations ($n = 56$) were .72 for the AI-4 and .69 for the AI-6. Criterion validity was also investigated using ROC analyses with the AI-4 and AI-6. A cutoff of ≥ 70 was utilized in all analyses, which included controls and clinical groups (ADHD, Emotional/Behavioral Disorder, Learning Disability,

Speech-Language Impairment, and Intellectual Disability). The AI-4 and AI-6 had AUC values in the good to excellent range (.82–.93) when classifying those with ASD vs. controls. AUC values were also good (.88–.89) when classifying ASD vs all clinical groups combined. However, singular diagnostic classifications were variable. AUCs in the poor range were found when classifying ASD vs. Emotional and Behavior Disorders and vs. Speech/Language Impairments (AUCs = .67–.69). However, ASD vs. ADHD was in the fair range (.72–.73), and ASD vs. Learning Disabilities was in the good range (.83–.84).

Finally, the author reported mean AI-4/AI-6 standard scores for the non-ASD groups investigated. The control mean for both the AI-4/6 index scores was ≤ 55 (Unlikely range). Regarding clinical groups, results were variable. Those with Intellectual Disability had AI-4/6 scores in the Very Likely range ($\bar{X}_{\text{standard scores}} = 87$ –89). With one exception (AI-6 in Learning Disabilities), mean scores for all other clinical groups were in the Likely range ($\bar{X}_{\text{standard scores}} = 57$ –64). However, all means were below the cutoff of 70 that was used in ROC analyses (Gilliam, 2014).

Why Additional Research is Important

The GARS-3 is considered a Level 2 measure (designed for use in diagnostic procedures where clinical groups are compared). The above research suggests that the GARS-3 has promising utility for differentiation between clinical groups, particularly using a cutoff of 70. However, additional research outside of the standardization sample is necessary. Reliance on the standardization sample information may over-estimate the utility of the GARS-3 in clinical groups (e.g., Ashwood et al., 2016; Camodeca, 2019; Camodeca et al., 2020; Camodeca & Walcott, 2021). Methodological challenges common to standardization samples may impact the generalizability of results. For the GARS-3, several methodological limitations exist.

First, ASD diagnosis (as well as other diagnostic classifications for non-ASD clinical groups) was not established as part of participation in the standardization research, and the degree to which a gold-standard measure of ASD diagnosis was utilized for diagnosis is unclear (Gilliam, 2014). Research indicates that use of gold-standard assessment measures in ASD diagnosis is highly variable overall, and is influenced by training and orientation (Aiello et al., 2017; Benson et al., 2019; Cook et al., 2017; Jensen-Doss & Hawley, 2010; Williams et al., 2009). However, the GARS-3 manual does not provide any information regarding the occupation or training of ASD diagnosticians (Gilliam, 2014). Relatedly, while high correlations were observed with the GARS-3 and the ADOS (Lord et al., 1999) in criterion validity investigation, the participants utilized ($n = 56$) represented only 3.01% of the overall sample. Additionally,

the field currently utilizes the ADOS-2 (Lord et al., 2012), a revision of the original ADOS that was utilized in Gilliam's (2014) research. For the ADOS-2 modules utilized in the current study, the manual indicates the revisions to the original ADOS do not fundamentally change the administration or coding requirements (Lord et al., 2012). However, comparison scores, which consider the child's age, have been provided for Modules 1, 2, and 3 of the ADOS-2; these comparison scores were not available in the original ADOS. Age being considered in comparison score conversion would appear to provide a purer and more normative analysis of ASD severity, as z-scores calculated by the GARS-3 author would be sample-dependent.

Second, research suggests comorbid disorders are common and often observed in the majority of each of the non-ASD clinical groups included in the GARS-3 investigations (Crisci et al., 2021; Efron et al., 2016; Gilliam, 2014; Reale et al., 2017; Stevens et al., 2016). In the standardization sample, comorbidity was represented in the ASD group, but the non-ASD clinical groups were comprised of children with single diagnoses only. This selection method may have improved the measure's ability to discriminate between groups due to their less complicated presentations (Manohar et al., 2018). Third, no comparisons were conducted regarding sex differences. Research suggests that while the core symptoms of ASD are present across sexes, there may be differences in presentation across sexes (Ramsey et al., 2018; Young et al., 2018). However, a recent review indicated that measurement invariance is observed across sexes for many measures (Lai & Szatmari, 2020), and many published ASD measures (e.g., Social Communication Questionnaire, Asperger Syndrome Diagnostic Scale, and Autism Spectrum Rating Scales) use the same norms for males and females. The use of the same norms across sexes for these similar, published, and frequently-utilized measures suggests that no sex differences would be observed for the GARS-3 as well, although ideally empirical support for this invariance would be obtained. Finally, there was limited investigation of the subscales regarding mean differences. The slightly higher AI-4/6 means observed for the ADHD sample vs. controls may reflect small elevations on all subscales, or higher scores on one or two subscales only. For example, those with ADHD have more social difficulties and problems with self-regulation than controls, meaning that the Social Interaction and Emotional Response subscales of the GARS-3 may be particularly elevated in that group (Bora & Pantelis, 2016; Goldstein & Naglieri, 2013; Salley et al., 2015).

Purpose and Hypotheses

There is a need for research investigating the GARS-3 criterion validity in a sample utilizing gold-standard measures as part of the diagnostic process. Analyses for this study

(*t*-tests/ANOVAs and ROC curves) were selected to replicate the standardization sample research. Additionally, Logistic Regression (LR) was added as use of both LR and ROC analyses is considered good practice for evaluating questionnaire criterion validity with binary outcomes (Youngstrom, 2014). For Hypotheses 1–3, analyses with all subscales and both the AI-4/6 scores were planned in order to provide an overview of the measure. In some prior research with ASD questionnaires, subscales have demonstrated more clinical utility (e.g., larger mean differences, higher AUC values, and larger Odds Ratios) than overall scores (Camodeca & Walcott, 2021; Camodeca, 2019). Thus, while Hypotheses 4 and 5 (ROC/LR) were primarily for the investigation of the AI-4/AI-6 scores, the author planned to investigate any promising subscales as determined by Hypotheses 2–3 with ROC/LR as well.

Hypothesis 1 Sex would be not be associated with GARS-3 Index score differences. By extension, it was hypothesized that the subscales would also not demonstrate sex differences. Gilliam (2014) indicated that correlations with age were weak; it is hypothesized that correlations in the current study will be similar. These hypotheses were tested first as sex and age could be utilized as control variables in further analyses.

Hypothesis 2 Mean Index and subscale scores of the GARS-3 would be higher in the ASD sample compared to the non-ASD clinical group.

Hypothesis 3 Correlations among AI-4/6 index scores and ADOS-2 scores would be similar to those observed in the standardization sample.

Hypothesis 4 ROC curves for the AI-4 and AI-6 would demonstrate an AUC in the fair range ($AUC = .70-.79$) in ASD vs. non-ASD comparisons.

Hypothesis 5 Effect sizes were not reported by the GARS-3 author. However, the mean AI-4/6 scores for the Non-ASD clinical groups were at least one and usually two standard deviations below the mean for the ASD group (using the normative data of $\bar{X} = 100$, $SD = 15$). With these numbers, a large effect size was estimated and thus moderate (5.0) to high (10.0+) Odds Ratios were expected.

Methods

Participants

Participants were 186 children aged 3–20 ($\bar{X}_{age} = 8.98$, $SD = 3.92$); $\bar{X}_{FSIQ} = 82.61$; $SD = 19.78$) referred for

evaluation with a question of ASD diagnosis. Referrals came from parents, special education personnel, mental health care providers (e.g., psychiatrists, psychologists, counselors, or in-home service agencies), pediatricians, or specialist medical practitioners (e.g., neurology, endocrinology, genetics). All children were referred for the purpose of diagnostic clarification and identification of strengths/weaknesses to inform appropriate treatment strategies. No children were referred due to participation in clinical trials, to document pre-operative functioning, or as part of ongoing medical or psychiatric treatment monitoring. Participants were classified into the ASD group (AUT; $n = 87$) if they received an ASD diagnosis as a result of their evaluation. All others were classified into a non-ASD group (NOT; $n = 99$). Level of ASD diagnosis was specified for 83.9% of the AUT group, and Language or Intellectual Impairment were specified for 88.5%. Of these, for Social Communication, 53.4% were Level 1, 43.8% were Level 2, and 2.7% were Level 3. For Restrictive and Repetitive Behaviors, 68.5% were Level 1, 28.8% were Level 2, and 2.7% were Level 3. A minority (19.5% and 13.0%) had a language impairment, and intellectual impairment, respectively. Detailed information regarding participants is provided in Table 1 and Supplementary Tables 1 and 2.

Materials

Demographics

Age, gender, race, handedness, and diagnoses had been entered into the dataset via review of parent-completed background information forms and clinician-completed assessment reports.

Gilliam Autism Rating Scale-3rd Edition (GARS-3)

Much of the information, including psychometric properties, regarding the GARS-3 was described in detail above. However, reliability information is provided here. Across all age groups, internal consistency reliability is very high for the Index scores ($\alpha \geq .91-.96$), and high to very high ($\alpha \geq .80-.96$) for all index scores except Maladaptive Speech—for this scale, some age groups were in the good range ($\alpha = .71-.85$). One- or two-week test-retest reliability was high for most scales/Indexes ($r = .80-.91$); Maladaptive Speech acceptable ($r = .76$).

Full Scale Intelligence Quotient (FSIQ)

FSIQs were obtained via standardized IQ tests as part of the assessment process. The Wechsler Intelligence Scale for Children-5th Edition (WISC-V; Wechsler, 2014) ($n = 102$) was utilized most frequently, followed by the Wechsler

Primary and Preschool Scale of Intelligence—4th Edition (WPPSI-IV) ($n = 37$) (Wechsler, 2012). Full Scale IQs were not available for 8.1% of cases ($n = 15$). All IQ tests used in the evaluations are standardized intellectual measures, and their reliability and validity is well-established (Reynolds & Kamphaus, 2003; Riccio et al., 2010; Sattler, 2018).

Autism Diagnostic Observation Schedule-2nd Edition (ADOS-2; Lord et al., 2012)

The ADOS-2 is a standardized, semi-structured observational assessment for ASD symptomatology and is considered one of the gold-standard assessments for ASD diagnosis. Examiners administering the ADOS-2 select from five Modules (1–4 and Toddler) based on the child's age and expressive language skills. Current study participants were administered Modules 1–4. Each module utilizes an algorithm consisting of codes of 0, 1, or 2 for various ASD symptoms (e.g., eye contact, verbal and non-verbal reciprocity gestures, rigidity, and sensory-seeking/aversions). Codes are summed and the total is compared to a Module-specific ASD cutoff score. Modules 1–3 provide an age-based ASD comparison conversion (1–10, higher scores associated with more symptomatology) for the raw score total (Lord et al., 2012). Inter-rater reliability is high (Kamp-Becker et al., 2018). Its validity has been well-established with children and adults with varying demographic characteristics (e.g., age, geographic location, comorbid diagnoses, and symptom severity) (Chojnicka & Pisula, 2017; Fusar-Poli et al., 2017; Langmann et al., 2017; Medda et al., 2019; Zander et al., 2016).

Procedure

Archival data were collected from an existing dataset containing results of comprehensive psychological/neuropsychological evaluations conducted between October 2014 and March 2020 at an outpatient community clinic in the eastern United States. Three licensed Psychologists (one Ph.D. and two Psy.D.) contributed data to the site's database. As the site specializes in ASD diagnosis, all clinicians are required to demonstrate clinical competence in the ADOS-2. This includes having prior ASD-assessment-focused clinical training, supervision of both administration and scoring from a clinical supervisor trained in the ADOS-2 on site, and completion of site-based trainings on the ADOS-2. As these data were archival, informed consent was not required. The study was conducted with full approval from the Institutional Review Board of the author.

The site has a uniform process for ASD assessment. All assessments consisted of an intake interview with the parent or guardian and the child or young adult being evaluated (examinee). Following the intake, standardized

Table 1 Demographic characteristics of sample and subgroups

Variable	Entire sample (<i>n</i> = 186)	Autism (AUT) (<i>n</i> = 87)	Not autism (NOT) (<i>n</i> = 99)	Significance test
Age	8.98 (3.92)	9.68 (4.09)	8.36 (3.70)	$t = -2.30, p = .02$
IQ				
All FSIQ (<i>n</i> = 171)	82.61 (19.78)	85.04 (20.21)	80.62 (19.29)	$t = -1.46, p = .15$
WISC-V (<i>n</i> = 102)	82.61 (19.45)	86.98 (19.72)	80.91 (18.97)	$t = -1.58, p = .12$
WPPSI-IV (<i>n</i> = 37)	81.41 (19.63)	78.28 (22.42)	84.37 (16.66)	$t = 0.94, p = .35$
Ethnicity (<i>n</i> = 183)				$\chi^2 = 1.67, p = .43$
Caucasian	84.9% (158)	87.4% (76)	82.8% (82)	
African American	4.8% (9)	4.6% (4)	5.1% (5)	
Other/Mixed	8.6% (16)	5.7% (5)	11.1% (11)	
Sex				$\chi^2 = 2.12, p = .15$
Male	73.1% (136)	78.2% (68)	68.7% (68)	
Female	26.9% (50)	21.8% (19)	31.3% (31)	
Handedness (<i>n</i> = 181)				$\chi^2 = 3.15, p = .21$
Right	77.4% (144)	73.6% (64)	80.8% (80)	
Left	15.1% (28)	19.5% (16)	11.1% (11)	
Ambidextrous	4.8% (9)	3.4% (3)	6.1% (6)	
ADOS-2 Module				$\chi^2 = 4.83, p = .19$
1	8.1% (15)	6.9% (6)	9.1% (9)	
2	22.6% (42)	16.1% (14)	28.3% (28)	
3	51.6% (96)	56.3% (49)	47.5% (47)	
4	17.7% (33)	20.7% (18)	15.2% (15)	
Diagnosis/Comorbidities				
Autism	46.8% (87)	100% (87)	—	
ADHD	68.3% (121)	65.5% (57)	70.7% (70)	—
Anxiety ^a	34.9% (65)	21.9% (20)	45.4% (45)	—
Mood disorders ^b	16.7% (31)	10.1% (9)	22.2% (22)	—
Intellectual disability ^c	10.3% (19)	8.1% (7)	12.1% (12)	—
Behavioral disorder	7.7% (14)	3.4% (3)	11.1% (11)	
Language impairment	3.8% (7)	1.1% (1)	6.1% (6)	—
Learning disabilities ^d	4.4% (6)	3.4% (3)	5.1% (5)	—
Other ^e	50.5% (94)	33.3% (29)	65.7% (65)	—

^aFive children had multiple anxiety disorders; 1 AUT, 4 NOT^bThree children had multiple mood disorders; 1 AUT, 2 NOT^cReflects children whose assessment reports included DSM-5 Fcodes for intellectual disabilities^dFour children had multiple learning disabilities; 1 AUT, 3 NOT^eEleven children had multiple Other disorders; 2 AUT, 9 NOT

assessment measures that were required to evaluate the diagnoses under investigation were provided to the examinee, and questionnaires were provided to their parent or guardian. The evaluations were tailored to the diagnoses under investigation, but generally included an assessment of cognitive functioning, attention, memory, and executive functioning, as well as broad-band and diagnosis-specific rating scales (e.g., ADHD, anxiety, depression). Examinees were also administered the ADOS-2 and at least one ASD questionnaire to evaluate ASD diagnosis. As the evaluations were comprehensive and not limited to ASD

symptoms only, all examinees received a diagnosis/explanation for current symptomatology after the assessment. Regarding ASD diagnosis in particular, ADOS-2 performance (i.e., algorithm scores and cutoffs), in conjunction with DSM-5 criteria, was given primary consideration. As questionnaires are not considered gold-standard assessments for ASD, the limitations of ASD questionnaires are known by the clinicians at the site, and the GARS-3 was given in the context of other ASD questionnaires, the GARS-3 was not a primary method utilized for ASD diagnosis.

Results

Preliminary Analyses

Sample size was sufficient for $\geq 80\%$ power assuming a medium effect size for group comparisons. Sample size was also sufficient for LR and ROC analyses (Tabachnick & Fidell, 2012). Prior to analyses, variables of interest were examined for accuracy and outliers/skew. Due to the large number of tests conducted in this study, the critical p value was reduced to .01. Preliminary analyses indicated no differences in demographic characteristics between the AUT/NOT groups, although age was marginal at $p = .02$. No differences in proportion of children with or without a FSIQ score were found between AUT/NOT groups, $X^2 = 2.59$, $p = .11$. No differences in any GARS-3 means were observed between children with or without a FSIQ, $t_s < |1.61|$, $p_s > .12$. FSIQ was significantly correlated with the Cognitive Style and Maladaptive Speech subscales, $r = .33$, $p < .001$; $r = -.27$, $p < .001$, respectively. For all other scores, $r_s \leq |.13|$, $p_s \geq .09$.

Hypothesis 1: Sex Differences and Correlations with Age

T-tests indicated no differences in GARS-3 scores based on sex, $t_s < |1.60|$, $p_s > .11$, although Social Interaction was marginal, $t = 2.13$, $p = .03$ (see Table 2). The hypothesis that no sex differences would be observed was supported.

Correlations with age are provided in Table 3. The GARS-3 author did not report p values associated with r s in age/score correlations. The p -value of r is irrelevant in Fisher- r -to- z comparisons. However, p -values are reported for the current sample because if significant correlations with age were found, age would be entered as a control variable. Age demonstrated weak negative correlations with both Restrictive/Repetitive Behavior and Emotional Response. Fisher- r -to- z comparisons indicated that the r for Social Interaction was higher in the current study than the standardization sample, $Z = 2.73$, $p = .01$. However, the correlation was still weak ($r = .17$) and non-significant. Due to the difference in r for the Social Interaction subscale, the Fisher- r -to- z comparisons do not entirely support the hypothesis that all age/score correlations observed in the current study are similar to Gilliam (2014). However, for practical purposes, the association between age and scores is minimal.

Table 2 T-tests with GARS-3 scales

Scale	Male ($n = 136$) \bar{X} (SD)	Female ($n = 50$) \bar{X} (SD)	t	p	Cohen's d
Restrictive/repetitive behaviors	7.74 (2.78)	7.46 (2.45)	0.08	.54	−0.10
Social Interaction	7.19 (2.62)	8.14 (2.87)	2.13	.03	0.35
Communication	6.93 (2.70)	7.62 (2.95)	1.50	.14	0.25
Emotional response	10.21 (3.20)	10.06 (3.30)	−0.29	.77	−0.05
Cognitive style	10.28 (2.39)	9.63 (2.53)	−1.60	.11	−0.27
Maladaptive speech	8.35 (2.73)	7.71 (2.31)	−1.45	.15	−0.24
GARS-3 Index-4 scales	86.57 (14.79)	88.54 (14.89)	0.80	.42	0.13
GARS-3 Index-6 scales	86.57 (17.05)	86.98 (15.62)	0.15	.89	0.02
Scale	AUT ($n = 87$) \bar{X} (SD)	NOT ($n = 99$) \bar{X} (SD)	t	p	Cohen's d
Restrictive/repetitive behaviors	7.72 (2.92)	7.61 (2.49)	−0.30	.77	−0.04
Social Interaction	7.89 (2.68)	7.06 (2.70)	−2.08	.04	−0.31
Communication	7.32 (2.62)	6.94 (2.91)	−0.04	.35	−0.14
Emotional response	9.86 (3.43)	10.45 (3.00)	1.27	.21	0.19
Cognitive style	10.43 (2.60)	9.82 (2.26)	−1.66	.10	−0.25
Maladaptive speech	8.14 (2.57)	8.22 (2.71)	0.20	.84	0.03
GARS-3 Index-4 scales	87.78 (15.19)	86.51 (14.51)	−0.59	.56	−0.09
GARS-3 Index-6 scales	87.56 (14.48)	85.89 (14.91)	−0.68	.50	−0.10

Table 3 Fisher *r*-to-*z* comparisons for age and ADOS-2 score correlations

Scale	Age		
	Gilliam (2014) <i>r</i> ^a	Current study <i>r</i>	<i>Z</i>
Restrictive/repetitive behaviors	−.15	−.19*	−0.53
Social Interaction	−.04	.17	−2.73*
Communication	−.02	−.04	−0.26
Emotional response	−.16	−.24**	−1.08
Cognitive style	.10	.12	0.26
Maladaptive speech	.01	−.17	−2.34
GARS-3 Index-4 scales	–	−.05	–
GARS-3 Index-6 scales	–	−.11	–
Scale	ADOS-2 Scores		
	Gilliam (2014) <i>r</i> ^a	Current study <i>r</i> ^b	<i>Z</i> ^b
GARS-3 Index-4 scales	.72	−.02/.08	5.81**/3.62**
GARS-3 Index-6 scales	.69	−.03/.12	5.48**/3.18**
Social Interaction	–	−.04/.37	–

p* < .01; *p* < .001^aSignificance levels were not indicated in Gilliam (2014)^bAnalysis with ADOS-2 Comparison Score, Modules 1–3/analyses with Social Communication and Social Interaction Total, Module 4

Hypothesis 2: Mean Differences for AUT/NOT Groups

T-tests indicated no significant differences in any GARS-3 scores when comparing AUT/NOT groups (see Table 3). ANCOVAs were conducted with age as a covariate for Restrictive/Repetitive Behavior and Emotional Response (see Table 4). For both scales, age was a significant covariate, but diagnostic group remained non-significant. ANCOVAs were conducted with FSIQ as a covariate for Cognitive Style and Maladaptive Speech. For both scales, FSIQ was a significant covariate, but diagnostic group remained non-significant. The hypothesis was not supported.

Hypothesis 3: Correlations with ADOS-2 Scores

As mentioned above, a more standardized method of measuring ASD symptomatology for correlations was desired in the current study vis-a-vis. Gilliam (2014). Additionally, as the current study utilized the ADOS-2, it was not possible to compute a communication/social interaction raw score total for all modules; furthermore, with changes between the ADOS and ADOS-2, it was uncertain if the items used to compute a communication/social interaction raw score total would be the same as Gilliam (2014). Thus the ADOS-2 comparison score (range = 1–10, with higher scores indicating more symptomatology) was utilized for correlations with the AI-4 and AI-6 for children who received Modules 1–3. There is no ADOS-2 comparison score for Module 4; these 4 participants were analyzed separately, using the

Table 4 ANCOVAs with GARS-3 scores

Scales	AUT \bar{X} (SE)	NOT \bar{X} (SE)	<i>F</i> value	<i>p</i> value	Cohen's <i>d</i> ^a
Restrictive/repetitive behavior ¹	7.82 (.29)	7.52 (.27)	0.57	.45	0.11
Emotional response ¹	9.99 (.34)	10.34 (.32)	0.57	.45	0.11
Cognitive style ²	10.43 (.26)	9.98 (.24)	1.63	.20	0.20
Maladaptive speech ²	8.29 (.29)	8.26 (.27)	0.01	.94	0.00

^aEta squared values for AUT/NOT were converted to Cohen's *d* for ease of comparison with prior analyses¹Control variable: age²Control variable: FSIQ

Table 5 Logistic regressions for Index Scores and Social Interaction

Scale	% Correct classification	Wald	<i>p</i> value	Odds ratio	95% CI
Autism Index 4	51.6%	0.35	.56	1.01	.99–1.03
Autism Index 6	54.9%	0.47	.49	1.01	.99–1.02
Social Interaction	57.5%	4.21	.04	1.12	1.01–1.25

Table 6 ROC curve analyses

Scale	AUC	Std. error	Significance	95% CI
Autism Index 4	.54	.04	.35	.46–.62
Autism Index 6	.54	.04	.35	.46–.62
Social Interaction	.59	.04	.04	.50–.67

raw Social Communication/Social Interaction Total score. Results of these analyses are presented in Table 3. Correlations were non-significant for both the AI-4 and AI-6 score across all Modules, $r_s < |.12|$, $p_s > .52$. All correlations were significantly lower than what was reported in Gilliam (2014), $Z_s > 3.18$, $p_s < .001$. The hypothesis was not supported. Exploratory analyses were conducted with the Social Interaction scale, although there were no comparison r_s from Gilliam (2014). The correlation for Modules 1–3 was weak and non-significant, and the correlation with Module 4 was moderate but marginal, $r = .37$, $p = .04$.

Hypothesis 4: Logistic Regressions

Given the marginal results observed in *t*-test analyses, the Social Interaction subscale was investigated in Logistic Regression along with the AI-4 and AI-6 (see Table 5). LR with the AI-4/6 correctly classified 51.6% and 54.9% of the sample, respectively, with $p_s \geq .49$. Both ORs 1.01, much smaller than expected based on Gilliam (2014), and the lower CI extended to .99 for both scales as well. Social Interaction was marginally significant in LR, OR 1.12[95%CI 1.01–1.25] $p = .04$, correctly classifying 57.5% of the sample. Again, however, the OR was small. The hypothesis was not supported.

Hypothesis 5: Receiver Operating Characteristic Analyses

ROC analyses were conducted with the AI-4, AI-6, and the Social Interaction subscale. Results of these analyses are presented in Table 6. Again, neither the AI-4 or AI-6 were significant. AUC was below the poor range; while the upper bound of the 95% CI extended into the poor range, the lower bound of the 95%CI extended below .50, where it would

predict the negative state. Social Interaction was again marginally significant, but the AUC was below the poor range. While the upper bound of the 95% CI extended into the poor range, the lower bound extended to .50, chance prediction. The hypothesis was not supported.

Classification statistics were investigated for the AI-4, AI-6, and Social Interaction (see Table 7). Initially, the author-specified cutoffs (55, 71, 101) were investigated. However, due to the large difference between the 71 and 101 classification statistics for both the AI-4 and AI-6, two middle scores were examined—88 and 95. Eighty-eight was selected as it was the whole number closest to the ASD group mean; 95 was selected as it was the approximate midpoint between 88 and 101.

The AI-4 classified the most children correctly with a cutoff score of 101 (55.91%); correct classification ranged from 46.24 to 54.30% with lower cutoffs. Sensitivity was in the excellent or better range at a cutoff of 55 and 71 (Likely and Very Likely—Level 2), but the false positive rates were very high (87.88% or higher). Both the True Positives (Sensitivity) and False Positives decreased the most between the 71 and 88 cutoffs; however, specificity (true negatives) did not increase to the excellent range until a cutoff of 101. The False Negative rate was $> 44.37\%$ considering all cutoff points, although the false positive rate decreased to 15.15% At the 101 cutoff. The highest predictive values (both positive and negative) occurred at the 101 cutoff and were better than chance, but only slightly. If a child scored at or above the 55/“Likely” cutoff, there was only a 46.45% that a child had ASD. There was no point that maximized both sensitivity and specificity, and unlike the standardization sample, the AI-4 did not appear much better at ruling out ASD with a low score.

Similarly, the AI-6 classified the most children correctly with a cutoff score of 101 (54.30%); correct classification ranged from 44.62 to 53.23% with lower cutoffs. Sensitivity was in the excellent or better range at a cutoff of 55 (Likely) and 71 (Very Likely—Level 2), but the false positive rates were very high (82.83% or higher). Both the True Positives (Sensitivity) and False Positives decreased the most between the 71 and 88 cutoffs; however, the False positive rate was still over 1/3 (36.36%) at a cutoff of 95 (between Very Likely Level 2 and Very Likely Level 3). Specificity (true negatives) was in the good (but very near excellent) at a cutoff of 101. The False Negative rate was $> 45.14\%$ considering all cutoff points, and decreased the most (by 35 percentage points) between the 55 and 71 cutoffs before leveling off. The highest predictive values (both positive and negative) occurred at the 101 cutoff and were better than chance, but only slightly. If a child scored at or above the 55/“Likely” cutoff, there was only a 45.56% chance that the child had ASD, and at the 101 cutoff, there was only a 52.38% chance the child had ASD. As with the AI-4, there was no point

Table 7 Classification statistics for cutoff values: Autism Index-4/6 and Social Interaction

	Autism Index-4 ^a				
	55 ^{1,2}	71 ³	88 ⁴	95	101 ⁴
Overall correct %	46.24	46.77	54.30	53.76	55.91
Sensitivity (true positive) %	97.70	86.21	55.17	40.23	22.99
Specificity (true negative) %	1.01	12.12	53.54	65.66	84.85
False positive rate %	98.99	87.88	46.46	34.34	15.15
False negative rate %	66.67	50.00	42.39	44.44	44.37
Positive predictive value	46.45	46.30	51.06	50.72	57.14
Negative predictive value	33.33	50.00	57.61	55.56	55.63
	Autism Index-6 ^a				
	55 ^{1,2}	71 ³	88 ⁴	95	101 ⁴
Overall correct %	44.62	47.31	53.23	52.69	54.30
Sensitivity (true positive) %	94.25	81.61	52.87	40.23	25.29
Specificity (true negative) %	1.01	17.17	53.54	63.64	79.80
False positive rate %	98.99	82.83	46.46	36.36	20.20
False negative rate %	83.33	48.48	43.62	45.22	45.14
Positive predictive value	45.56	46.41	50.00	49.30	52.38
Negative predictive value	16.67	51.52	56.38	54.78	54.86
	Social Interaction ^b				
	4	5	8	10	11
Overall correct %	49.46	52.69	54.84	55.91	56.99
Sensitivity (true positive) %	96.55	87.36	52.87	28.74	19.54
Specificity (true negative) %	8.08	22.22	56.57	79.80	89.90
False positive rate %	91.92	77.78	43.43	20.20	10.10
False negative rate %	27.27	33.33	42.27	43.97	44.03
Positive predictive value	48.00	49.67	51.69	55.56	62.96
Negative predictive value	72.73	66.67	57.73	56.03	55.97

^aStandard Scores ($\bar{X} = 100$, $SD = 15$)

^bScaled Scores ($\bar{X} = 10$, $SD = 3$)

¹≤ 54 (unlikely) had the same values as the 55 cutoff

²Probable (55–70; Level 1)

³Very likely (71–100; Level 2)

⁴Very likely (101+, Level 3)

that maximized both sensitivity and specificity, and unlike the standardization sample, the AI-4 did not appear better at ruling out ASD with a low score.

There were no GARS-3-author-provided cutoffs for the Social Interaction scale. The possible range of scores is between 3 (1st percentile)–14 (91st percentile). The author of this paper matched the percentile ranks of the scores utilized for the AI-4/6 cutoff scores as closely as possible. There was no option for the 55/below the 1st percentile cutoff as the first percentile scaled score of 3 was the lowest score; thus a cutoff of 4 (2nd percentile) was used. A scaled score of 5 (5th percentile) was used in place of the 71/3rd percentile score. The scaled score mean of 10 was equal to the 50th percentile, and the 101 cutoff was just above that

(53rd percentile). As these analyses were exploratory, it was decided to investigate both the 10 and 11 (63rd percentile) cutoff points. A cutoff of 12 (75th percentile) was included as a) the prior indexes had each had 5 cutoff scores examined, and b) examination of frequency distributions/standard deviations indicated few scores above 12.

The Social Interaction scale classified the most children correctly with a cutoff score of 11 (56.99%); correct classification ranged from 52.69 to 54.84% with lower cutoffs. Sensitivity (true positives) was in the excellent range at a cutoff of 5, but quickly dropped to poor with a cutoff of 8. The False Positive rates were high (> 43.43) with cutoffs of 5 or 8, reduced to ~ 20% with a cutoff of 10, and ~ 10% with a cutoff of 11. Specificity (true negatives) was in the good

(but very near excellent) range at a cutoff of 10, and in the excellent range at a cutoff of 11. The False Negative rate was relatively stable compared to the AI-4/6 scales, ranging from a low of 33.33% at a cutoff of 5 and 45.56% at a cutoff of 12. However, these percentages were still fairly high. The highest positive predictive value occurred at the 11 cutoff (62.96%), and the highest negative predictive value at the 10 cutoff. The positive predictive value at the 11 cutoff was higher than that observed with the AI-4/6. As with the AI-4 and the AI-6, there was no point that maximized both sensitivity and specificity. However, the True Negative vs the True Positive/False Positive rates suggest that the Social Interaction scale might be useful in ruling out ASD in cases of a low score. This pattern (better at ruling out ASD with low scores) was observed in the standardization sample(s) (Gilliam, 2014).

Exploratory Analyses Regarding the Sample

Exploratory analyses were conducted in an attempt to explain these unexpected results. In particular, the AUT group mean appeared significantly lower, and the NOT group mean significantly higher, than the respective group means reported by Gilliam (2014). The discrepancy was investigated via 1-sample *t*-tests. Results indicated that both the AI-4 and AI-6 means of the AUT group were lower than Gilliam's (2014) ASD group (test value = 100), $t_s \geq 16.64$, $ps \leq .001$.

There was no single mean to select for a Non-Autism clinical group, as means were reported separately for Intellectual disabilities, ADHD, Emotional/Behavioral Disorders, Learning Disabilities, and Speech/Language impairment. All these comparison groups were represented in the current sample, but ADHD had the highest incidence. Thus two sets of 1-sample *t*-tests were run—the first set compared the sample against Gilliam's (2014) ADHD group ($\bar{X}_{AI-4} = 61$; $\bar{X}_{AI-6} = 55$) and the second set compared the sample against the average of the means for all comparison groups ($\bar{X}_{AI-4} = 66.60$; $\bar{X}_{AI-6} = 62.40$). In all four analyses, the current NOT sample means were significantly higher than that of the non-ASD clinical group Gilliam (2014), $t_s \geq 14.06$, all $ps \leq .001$. The exploratory hypothesis was supported.

Discussion

Summary

The current study attempted to improve upon the methodology utilized in Gilliam (2014) in establishing the standardization sample for the GARS-3, and to investigate the utility of the GARS-3 in diagnosis of a sample with comorbidities.

Hypotheses regarding age and sex were mostly supported, but hypotheses regarding criterion validity were not.

Age Correlations and Sex Differences

Age

Gilliam (2014) reported small *r* values for age correlations in the standardization sample; it was hypothesized that correlations in the current study would be similar in value. This hypothesis was supported for all scales except Social Interaction, and in that case the correlation was weak and non-significant. Considering the *r*-value, however, perhaps a better hypothesis would have been to state that the correlations in this sample would also be weak according to the guidelines used in Gilliam's (2014) sample. Using that interpretation, the hypothesis is supported for all scales.

Although the *r*-values themselves were similar across studies, without any significance levels presented by the author (Gilliam, 2014) it is difficult to interpret the significant correlations with age and the non-significant Fisher-*r*-to-*z* transformations for Emotional Response and Restrictive/Repetitive Behaviors. Age was a significant covariate in both ANCOVAs, suggesting a small but potentially meaningful influence of age. There are several ASD measures for which age is considered for normative scoring, such as the Social Responsiveness Scale-2 and the Autism Spectrum Rating Scales, although those scales have broad age groups (Constantino & Gruber, 2012; Goldstein & Naglieri, 2013). It is possible that age is a relevant covariate in referred samples for behaviors on these scales—that is, perceived developmentally inappropriate behaviors in these areas are associated with increased referral for ASD assessment. This idea was supported in prior measure-based research in preschoolers referred for ASD assessment, as well as in a prior study examining referral for assessment based on child observation (Camodeca & Walcott, 2021; Rosenbaum & Gabrielsen, 2019). However, as the correlations were weak, the idea that age is relevant in clinical samples for some constructs should be investigated in future research.

Sex

There were no sex differences regarding any scales. Sex differences were not investigated by Gilliam (2014). There are several scales that do not differentiate between sexes regarding normative data or clinical cutoffs, and the current findings support this lack of differentiation. Others scales do have different norms for each sex. Potential differences in questions between measures or any referral bias (decreased referral of girls who do not demonstrate typical ASD behavior) should be investigated in future research.

Mean Differences, LRs, and ROCs

There were no mean differences between diagnostic groups on t-tests or ANCOVAs for any scales. Results of LR and AUC analyses were also non-significant. Sensitivity and specificity were below the values reported by the GARS-3 author, and could not be optimized at any cutpoint. These non-significant findings for the AI-4 and AI-6 are particularly important regarding the practical utility of the GARS-3, as these scores are associated with the probability of ASD diagnosis (Gilliam, 2014). Correlations with the ADOS-2 were significantly weaker compared to the GARS-3 sample (Gilliam, 2014). The current study's findings suggest the GARS-3 psychometric properties are below expectations for a measure intended for use in ASD assessment/diagnostic procedures (Youngstrom, 2014). While many scales can be utilized as screeners by ruling out ASD in cases of low scores, the GARS-3 AI-4/6 scores have a false negative rate of 67–83%, respectively, at the lowest cutoffs. The highest cutoffs investigated also demonstrated high (44–45%) false negative rates. It appears that the Social Interaction scale is somewhat more effective, in that scores of 5 or below are more likely to be Non-ASD diagnoses. These findings are consistent with the prior research conducted with the GARS-2, suggesting the revisions in the new edition were not substantial enough to improve the psychometric properties of the measure (Hampton & Strand, 2015; Norris & LeCavalier, 2010; Pandolfi et al., 2010).

The current study's sample composition provides several potential explanations for the poor diagnostic utility of the GARS-3 vis-à-vis Gilliam (2014). First, the most frequent diagnosis in the NOT group was ADHD, which can be difficult to differentiate from ASD (Manohar et al., 2018). This difficulty was illustrated in Gilliam's study, in which the AUC was in the good (as opposed to the excellent) range for ASD/ADHD group comparisons. However, the current study's AUC was non-significant and in the poor range, lower than Gilliam's standardization sample (2014). Additionally, in Gilliam's (2014) sample, the difficulty differentiating the ASD and ADHD groups appeared related to higher scores in the ADHD group. In the current study, the poor differentiation appeared related to both higher scores in the NOT group and lower scores in the AUT group. Thus difficulty with ADHD differentiation does not appear to explain the findings entirely.

A second explanation is the number and type of comorbidities in the current sample. This study's participants had an average of 2.3 diagnoses each, whereas the standardization sample had only one (small) comorbid sub-group with one comorbidity (Autism + ADHD). Research on comorbidities in both children and adults suggests that comorbid

ASD is associated with less differentiation between ASD/non-ASD clinical groups on both ASD and general psychopathology questionnaires (Dudas et al., 2017; Yamawaki et al., 2020). Another factor potentially increasing NOT group scores is the presence of emotional and behavior challenges, which is both common in referred samples and associated with higher scores on autism measures (Akmatov et al., 2021; Avni et al., 2018; Efron et al., 2016; Joshi et al., 2010; Magyar & Pandolfi, 2017; Rieske et al., 2015). In the NOT sample, 45.4% had an anxiety disorder, 18.5% had a mood disorder, and 65.7% had a behavioral disorder. In the GARS-3 standardization sample, poor range AUCs were found when comparing between the ASD and the Emotional/Behavioral Diagnosis group in particular. Additional research on the GARS-3 may help improve the psychometric properties in complex samples.

Third, it is possible the current study's participants demonstrated milder ASD symptomatology (mostly Level 1 and 2) compared to Gilliam's (2014) sample. Perhaps the majority of Gilliam's (2014) sample was diagnosed with Level 2 or 3 ASD, which increased the bar against which normative scores were created. Level of functioning was not specified in Gilliam's (2014) sample, making it difficult to judge this possibility. Regardless of the explanation, however, the GARS-3 is intended to differentiate between autism/non-ASD groups with functioning at all levels. Fourth, the current study was prospective, involving parents who did not know if their child had autism, while Gilliam's (2014) sample included previously diagnosed participants. It is possible that the assessment process in general educates parents regarding their child's symptoms (e.g., what is representative of ASD vs. non-ASD diagnoses), and could influence ratings of their child's behavior (Austin et al., 2019; Shepherd et al., 2018; Wigham et al., 2019). It would potentially be beneficial to develop measures in a sample of yet-to-be-diagnosed participants. Finally, as mentioned previously, the method of diagnosis in the standardization sample is unknown. It is possible that lack of a gold-standard measure for ASD in some or all of the standardization sample resulted in unreliable or invalid diagnosis, and thus poor between-group differentiation on the GARS-3.

Correlations with the ADOS-2

The correlations among the ADOS-2 and AI-4/6 scores were significantly weaker compared to the standardization sample. Given the findings in all other analyses, it is also probable that the poor psychometric properties of the GARS-3 in ASD diagnosis in complex samples are related to the significantly weaker correlations observed. The results may also be related to the different measure (ASD comparison

score for Modules 1–3, and raw Social Interaction and Communication total for Module 4) utilized in the current study. However, as mentioned above, the current study's measure provided a more standardized and replicable method of measuring ADOS-2 performance.

Exploratory Analyses for Social Interaction

The findings for Social Interaction were marginal in all analyses except for the Module 1–3 comparison score correlation, which was non-significant. Nonetheless, these findings were comparatively stronger compared to the AI-4/6. Two recent studies with the Autism Spectrum Rating Scales parent report 2–5 (Camodeca & Walcott, 2021) and 6–18 (Camodeca, 2019) have demonstrated findings similar to the current study. In both studies, the scales intended for diagnosis demonstrated weak or non-significant findings in analyses conducted to establish criterion validity. However, scales measuring peer interactions, social skills, and social-emotional reciprocity demonstrated stronger findings. A review of ASD questionnaires also indicated that socially-oriented scales demonstrated the best diagnostic utility in ASD (Hampton & Strand, 2015). The Social Interaction subscale likely emerged as a stronger scale in this study due to congruence with core ASD symptomatology (i.e., social communication weaknesses) (Dolan et al., 2016; Espelöer et al., 2021; Richey et al., 2014; Wagner et al., 2018). Thus, despite the marginal findings, the Social Interaction scale may be a useful starting point to improve upon the psychometric properties of the GARS-3.

Strengths and Limitations

The current study has many strengths. To the author's knowledge, this is the only study outside of the standardization sample that has investigated the psychometric properties of the GARS-3. A comprehensive assessment was conducted for all participants, and a gold-standard measure was utilized for ASD evaluation. The study had a large sample size. Limitations are related to the use of clinical and archival data. There was no ability to randomize measure presentation. The number of questionnaires completed by each parent for their child was variable, and depended on the clinician, age of the child, and what other diagnoses were under investigation. Additionally, parent variables that may have impacted responses, such as knowledge of ASD, interpretation of questions, and stress associated with parenting a child with special needs were not evaluated. However, the conditions under which the data were collected are typical of the diagnostic process.

Conclusions and Implications

The GARS-3 diagnostic scales (AI-4/6) do not demonstrate adequate criterion validity to utilize in diagnostic procedures. The Social Interaction scale demonstrated relatively stronger findings, and has the potential to be utilized to rule out ASD in cases of low scores. However, there are other published screeners available with better predictive utility than the Social Interaction scale (Hampton & Strand, 2015; Norris & Lecavalier, 2010). These findings underscore the conceptualization of ASD as a social disorder and the need for research on ASD questionnaires outside of standardization samples. As the GARS-3 is designed to be used prospectively, and the sample utilized is representative of clinically-referred children, additional modification of this measure would be required for it to be utilized in the assessment process.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10803-022-05483-5>.

Acknowledgments This study did not receive funding. Limited preliminary data on the ROCs were presented as part of a larger presentation at the Center for Autism and Related Disorders conference.

Declarations

Conflict of interest The authors have no conflicts of interest to declare.

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent Informed consent was obtained from all individual participants prior to participation in this study.

References

- Aiello, R., Ruble, L., & Esler, A. (2017). National study of school psychologists' use of evidence-based assessment in autism spectrum disorder. *Journal of Applied School Psychology*, 33(1), 67–88. <https://doi.org/10.1080/15377903.2016.1236307>
- Akmatov, M. K., Ermakova, T., & Bätzing, J. (2021). Psychiatric and nonpsychiatric comorbidities among children with ADHD: An exploratory analysis of nationwide claims data in Germany. *Journal of Attention Disorders*, 25(6), 874–884. <https://doi.org/10.1177/1087054719865779>
- Alsaedi, R. H., Carrington, S., & Watters, J. J. (2020). Behavioral and neuropsychological evaluation of executive functions in children with autism spectrum disorder in the Gulf region. *Brain Sciences*, 10(2), 120. <https://doi.org/10.3390/brainsci10020120>
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). American Psychiatric Publishing.
- Ashwood, K. L., Gillan, N., Horder, J., Hayward, H., Woodhouse, E., McEwen, F. S., Findon, J., Eklund, H., Spain, D., Wilson,

- C. E., Cadman, T., Young, S., Stoencheva, V., Murphy, C. M., Robertson, D., Charman, T., Bolton, P., Glaser, K., Asherson, P., Simonoff, E., & Murphy, D. G. (2016). Predicting the diagnosis of autism in adults using the autism-spectrum quotient (AQ) questionnaire. *Psychological Medicine*, 46(12), 2595–2604. <https://doi.org/10.1017/S0033291716001082>
- Austin, C. A., Gerstle, M., Baum, K. T., Bradley, A., LeJeune, B., Peugh, J., & Beebe, D. W. (2019). Evolution of parental knowledge and efficacy across the pediatric neuropsychological evaluation process. *Clinical Neuropsychologist*, 33(4), 743–759. <https://doi.org/10.1080/13854046.2018.1497206>
- Avni, E., Ben-Itzhak, E., & Zachor, D. A. (2018). The presence of comorbid ADHD and anxiety symptoms in autism spectrum disorder: Clinical presentation and predictors. *Frontiers in Psychiatry*, 9, 717–717. <https://doi.org/10.3389/fpsyg.2018.00717>
- Basiru, T., Adereti, I., Olanipekun, A., & Ravenscroft, S. (2021). Trend in age at first diagnosis of autism spectrum disorder (ASD): Analysis of the 2012–2019 national survey of children's health (NSCH) data. *Annals of Epidemiology*, 61, 20–20. <https://doi.org/10.1016/j.annepidem.2021.05.017>
- Benson, N. F., Floyd, R. G., Kranzler, J. H., Eckert, T. L., Fefer, S. A., & Morgan, G. B. (2019). Test use and assessment practices of school psychologists in the united states: Findings from the 2017 national survey. *Journal of School Psychology*, 72, 29–48. <https://doi.org/10.1016/j.jsp.2018.12.004>
- Bora, E., & Pantelis, C. (2016). Meta-analysis of social cognition in attention-deficit/hyperactivity disorder (ADHD): Comparison with healthy controls and autistic spectrum disorder. *Psychological Medicine*, 46(4), 699–716. <https://doi.org/10.1017/S0033291715002573>
- Camodeca, A. (2019). Description of criterion validity of the autism spectrum rating scales 6–18 parent report: Initial exploration in a large community sample. *Child Psychiatry & Human Development*, 50(6), 987–1001. <https://doi.org/10.1007/s10578-019-00899-0>
- Camodeca, A., Todd, K., & Croyle, J. (2019 online, 2020). Utility of the Asperger Syndrome Diagnostic Scale in the assessment of autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 50, 513–523. <https://doi.org/10.1007/s10803-019-04272-x>
- Camodeca, A., & Walcott, K. (2021). Criterion validity of the autism spectrum rating scales 2–5 parent report. *Research in Autism Spectrum Disorders*, 86, 101820. <https://doi.org/10.1016/j.rasd.2021.101820>
- Cardon, T., Wangsgard, N., & Dobson, N. (2019). Video modeling using classroom peers as models to increase social communication skills in children with ASD in an integrated preschool. *Education & Treatment of Children*, 42(4), 515–536. <https://doi.org/10.1353/etc.2019.0024>
- Charman, T., & Gotham, K. (2013). Measurement issues: Screening and diagnostic instruments for autism spectrum disorders—Lessons from research and practise. *Child and Adolescent Mental Health*, 18(1), 52–63. <https://doi.org/10.1111/j.1475-3588.2012.00664.x>
- Chojnicka, I., & Pisula, E. (2017). Adaptation and validation of the ADOS-2, Polish version. *Frontiers in Psychology*, 8, 1916–1916. <https://doi.org/10.3389/fpsyg.2017.01916>
- Constantino, J. N., & Gruber, C. P. (2012). *Social Responsiveness Scale—Second Edition (SRS-2)*. Western Psychological Services.
- Cook, J. R., Hausman, E. M., Jensen-Doss, A., & Hawley, K. M. (2017). Assessment practices of child clinicians: Results from a national survey. *Assessment*, 24(2), 210–221. <https://doi.org/10.1177/1073191115604353>
- Crisci, G., Caviola, S., Cardillo, R., & Mammarella, I. C. (2021). Executive functions in neurodevelopmental disorders: Comorbidity overlaps between attention deficit and hyperactivity disorder and specific learning disorders. *Frontiers in Human Neuroscience*, 15, 594234. <https://doi.org/10.3389/fnhum.2021.594234>
- Davidovitch, M., Slobodin, O., Weisskopf, M. G., & Rotem, R. S. (2020). Age-specific time trends in incidence rates of autism spectrum disorder following adaptation of DSM-5 and other ASD-related regulatory changes in Israel. *Autism Research*, 13(11), 1893–1901. <https://doi.org/10.1002/aur.2420>
- Dolan, B. K., Van Hecke, A. V., Carson, A. M., Karst, J. S., Stevens, S., Schohl, K. A., Potts, S., Kahne, J., Linneman, N., Rummel, R., & Hummel, E. (2016). Brief report: Assessment of intervention effects on in-vivo peer interactions in adolescents with autism spectrum disorder (ASD). *Journal of Autism and Developmental Disorders*, 46(6), 2251–2259. <https://doi.org/10.1007/s10803-016-2738-0>
- Dudas, R. B., Lovejoy, C., Cassidy, S., Allison, C., Smith, P., & Baron-Cohen, S. (2017). The overlap between autistic spectrum conditions and borderline personality disorder. *PLoS ONE*, 12(9), e0184447. <https://doi.org/10.1371/journal.pone.0184447>
- Efron, D., Bryson, H., Lycett, K., & Sciberras, E. (2016). Children referred for evaluation for ADHD: Comorbidity profiles and characteristics associated with a positive diagnosis. *Child: Care, Health & Development*, 42(5), 718–724. <https://doi.org/10.1111/cch.12364>
- Espelöer, J., Hellmich, M., Vogeley, K., & Falter-Wagner, C. M. (2021). Brief report: Social anxiety in autism spectrum disorder is based on deficits in social competence. *Journal of Autism and Developmental Disorders*, 51(1), 315–322. <https://doi.org/10.1007/s10803-020-04529-w>
- Fusar-Poli, L., Brondino, N., Rocchetti, M., Panisi, C., Provenzani, U., Damiani, S., & Politi, P. (2017). Diagnosing ASD in adults without ID: Accuracy of the ADOS-2 and the ADI-R. *Journal of Autism and Developmental Disorders*, 47(11), 3370–3379. <https://doi.org/10.1007/s10803-017-3258-2>
- Gilliam, J. (2006). *Gilliam autism rating scales* (2nd ed.). Pearson.
- Gilliam, J. (2014). *Gilliam autism rating scales* (3rd ed.). Pearson.
- Goldstein, S., & Naglieri, J. A. (2013). *Autism Spectrum Rating Scales (ASRS): Technical manual*. Multi-Health Systems, Inc.
- Hamad, A. F., Alessi-Severini, S., Mahmud, S. M., Brownell, M., & Kuo, I. F. (2019). Annual trends in prevalence and incidence of autism spectrum disorders in Manitoba preschoolers and toddlers: 2004–2015. *Canadian Journal of Public Health*, 110(4), 476–484. <https://doi.org/10.17269/s41997-019-00181-9>
- Hampton, J., & Strand, P. S. (2015). A review of level 2 parent-report instruments used to screen children aged 1.5–5 for autism: A meta-analytic update. *Journal of Autism and Developmental Disorders*, 45(8), 2519–2530. <https://doi.org/10.1007/s10803-015-2419-4>
- Jensen-Doss, A., & Hawley, K. M. (2010). Understanding barriers to evidence-based assessment: Clinician attitudes toward standardized assessment tools. *Journal of Clinical Child and Adolescent Psychology*, 39(6), 885–896. <https://doi.org/10.1080/15374416.2010.517169>
- Joshi, G., Petty, C., Wozniak, J., Henin, A., Fried, R., Galdo, M., Kotarski, M., Walls, S., & Biederman, J. (2010). The heavy burden of psychiatric comorbidity in youth with autism spectrum disorders: A large comparative study of a psychiatrically referred population. *Journal of Autism and Developmental Disorders*, 40(11), 1361–1370. <https://doi.org/10.1007/s10803-010-0996-9>
- Kamp-Becker, I., Albertowski, K., Becker, J., Ghahreman, M., Langmann, A., Mingeback, T., Poustka, L., Weber, L., Schmidt, H., Smidt, J., Stehr, T., Roessner, V., Kucharczyk, K., Wolff, N., & Stroth, S. (2018). Diagnostic accuracy of the ADOS and ADOS-2 in clinical practice. *European Child & Adolescent Psychiatry*, 27(9), 1193–1207. <https://doi.org/10.1007/s00787-018-1143-y>
- Knowland, V. C. P., Fletcher, F., Henderson, L., Walker, S., Norbury, C. F., & Gaskell, M. G. (2019). Sleep promotes phonological

- learning in children across language and autism spectra. *Journal of Speech, Language, and Hearing Research*, 62(12), 4235–4255. https://doi.org/10.1044/2019_JSLHR-S-19-0098
- Lai, M., & Szatmari, P. (2020). Sex and gender impacts on the behavioural presentation and recognition of autism. *Current Opinion in Psychiatry*, 33(2), 117–123. <https://doi.org/10.1097/YCO.0000000000000575>
- Langmann, A., Becker, J., Poustka, L., Becker, K., & Kamp-Becker, I. (2017). Diagnostic utility of the autism diagnostic observation schedule in a clinical sample of adolescents and adults. *Research in Autism Spectrum Disorders*, 34, 34–43. <https://doi.org/10.1016/j.rasd.2016.11.012>
- Lord, C., Rutter, M., DiLavore, P. C., & Risi, S. (1999). *Autism Diagnostic Observation Schedule*. Western Psychological Services.
- Lord, C., Rutter, M., DiLavore, P. C., Risi, S., Gotham, K., & Bishop, S. L. (2012). *ADOS-2. Autism diagnostic observation schedule 2nd edition manual (Part I): Modules 1–4*. Western Psychological Services.
- Lordo, D., Bertolin, M., Sudikoff, E., Keith, C., Braddock, B., & Kaufman, D. A. S. (2017). Parents perceive improvements in socio-emotional functioning in adolescents with ASD following social skills treatment. *Journal of Autism and Developmental Disorders*, 47(1), 203–214. <https://doi.org/10.1007/s10803-016-2969-0>
- Magyar, C. I., & Pandolfi, V. (2017). Utility of the CBCL DSM-oriented scales in assessing emotional disorders in youth with autism. *Research in Autism Spectrum Disorders*, 37, 11–20. <https://doi.org/10.1016/j.rasd.2017.01.009>
- Manohar, H., Kuppili, P. P., Kandasamy, P., Chandrasekaran, V., & Rajkumar, R. P. (2018). Implications of comorbid ADHD in ASD interventions and outcome: Results from a naturalistic follow up study from south India. *Asian Journal of Psychiatry*, 33, 68–73. <https://doi.org/10.1016/j.ajp.2018.03.009>
- Medda, J. E., Cholemkery, H., & Freitag, C. M. (2019). Sensitivity and specificity of the ADOS-2 algorithm in a large German sample. *Journal of Autism and Developmental Disorders*, 49(2), 750–761. <https://doi.org/10.1007/s10803-018-3750-3>
- Norris, M., & Lecavalier, L. (2010). Screening accuracy of level 2 autism spectrum disorder rating scales. *Autism*, 14(4), 263–284. <https://doi.org/10.1177/1362361309348071>
- Pandolfi, V., Magyar, C. I., & Dill, C. A. (2010). Constructs assessed by the GARS-2: Factor analysis of data from the standardization sample. *Journal of Autism and Developmental Disorders*, 40(9), 1118–1130. <https://doi.org/10.1007/s10803-010-0967-1>
- Pfeiffer, B., Piller, A., Slugg, L., & Shiu, C. (2018). Brief report: Reliability of the participation and sensory environment questionnaire: Home scales. *Journal of Autism and Developmental Disorders*, 48(7), 2567–2576. <https://doi.org/10.1007/s10803-018-3499-8>
- Ramsey, R. K., Nichols, L., Ludwig, N. N., Fein, D., Adamson, L. B., & Robins, D. L. (2018). Brief report: Sex differences in parental concerns for toddlers with autism risk. *Journal of Autism and Developmental Disorders*, 48(12), 4063–4069. <https://doi.org/10.1007/s10803-018-3583-0>
- Reale, L., Bartoli, B., Cartabia, M., Zanetti, M., Costantino, M. A., Canevini, M. P., Termine, C., Bonati, M., & Lombardy ADHD Group. (2017). Comorbidity prevalence and treatment outcome in children and adolescents with ADHD. *European Child & Adolescent Psychiatry*, 26(12), 1443–1457. <https://doi.org/10.1007/s00787-017-1005-z>
- Reynolds, C. R., & Kamphaus, R. W. (2003). *Handbook of psychological and educational assessment of children: Intelligence, aptitude, and achievement* (2nd ed.). Guilford Press.
- Riccio, C. A., Sullivan, J. R., & Cohen, M. J. (2010). *Neuropsychological assessment and intervention for childhood and adolescent disorders*. Wiley.
- Richey, J. A., Rittenberg, A., Hughes, L., Damiano, C. R., Sabatino, A., Miller, S., Hanna, E., Bodfish, J. W., & Dichter, G. S. (2014). Common and distinct neural features of social and non-social reward processing in autism and social anxiety disorder. *Social Cognitive and Affective Neuroscience*, 9(3), 367–377. <https://doi.org/10.1093/scan/nss146>
- Rieske, R. D., Matson, J. L., Beighley, J. S., Cervantes, P. E., Goldin, R. L., & Jang, J. (2015). Comorbid psychopathology rates in children diagnosed with autism spectrum disorders according to the DSM-IV-TR and the proposed DSM-5. *Developmental Neuropsychology*, 18(4), 218–223. <https://doi.org/10.3109/17518423.2013.790519>
- Rosenbaum, M., & Gabrielsen, T. P. (2019). Decision factors for community providers when referring very young children for autism evaluation. *Research in Autism Spectrum Disorders*, 57, 87–96. <https://doi.org/10.1016/j.rasd.2018.09.009>
- Salley, B., Gabrielli, J., Smith, C. M., & Braun, M. (2015). Do communication and social interaction skills differ across youth diagnosed with autism spectrum disorder, attention-deficit/hyperactivity disorder, or dual diagnosis? *Research in Autism Spectrum Disorders*, 20, 58–66. <https://doi.org/10.1016/j.rasd.2015.08.006>
- Sattler, J. (2018). *Assessment of children: Cognitive foundations and applications* (2nd ed.). Jerome Sattler Publications.
- Shepherd, D., Landon, J., Goedeke, S., Ty, K., & Csako, R. (2018). Parents' assessments of their child's autism-related interventions. *Research in Autism Spectrum Disorders*, 50, 1–10. <https://doi.org/10.1016/j.rasd.2018.02.005>
- Stevens, T., Peng, L., & Barnard-Brak, L. (2016). The comorbidity of ADHD in children diagnosed with autism spectrum disorder. *Research in Autism Spectrum Disorders*, 31, 11–18. <https://doi.org/10.1016/j.rasd.2016.07.003>
- Tabachnick, B. G., & Fidell, L. S. (2012). *Using multivariate statistics* (6th ed.). Pearson.
- Tse, C. Y. A., Pang, C. L., & Lee, P. H. (2018). Choosing an appropriate physical exercise to reduce stereotypic behavior in children with autism spectrum disorders: A non-randomized crossover study. *Journal of Autism and Developmental Disorders*, 48(5), 1666–1672. <https://doi.org/10.1007/s10803-017-3419-3>
- Wagner, J., Luyster, R. J., Moustapha, H., Tager-Flusberg, H., & Nelson, C. A. (2018). Differential attention to faces in infant siblings of children with autism spectrum disorder and associations with later social and language ability. *International Journal of Behavioral Development*, 42(1), 83–92. <https://doi.org/10.1177/0165025416673475>
- Wechsler, D. (2012). *Technical and interpretative manual of the Wechsler preschool and primary scale of intelligence 2nd edition*. PsychCorp.
- Wechsler, D. (2014). *Technical and interpretative manual of the Wechsler intelligence scale for children 5th edition*. PsychCorp.
- 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>
- Williams, M. E., Williams, M. E., Atkins, M., Atkins, M., Soles, T., & Soles, T. (2009). Assessment of autism in community settings: Discrepancies in classification. *Journal of Autism and Developmental Disorders*, 39(4), 660–669. <https://doi.org/10.1007/s10803-008-0668-1>
- Yamawaki, K., Ishitsuka, K., Suyama, S., Suzumura, S., Yamashita, H., & Kanba, S. (2020). Clinical characteristics of boys with comorbid autism spectrum disorder and attention deficit/hyperactivity disorder. *Pediatrics International*, 62(2), 151–157. <https://doi.org/10.1111/ped.14105>

- Young, H., Oreve, M. J., & Speranza, M. (2018). Clinical characteristics and problems diagnosing autism spectrum disorder in girls. *Archives De Pédiatrie: Organe Officiel De La Société Française De Pédiatrie*, 25(6), 399–403. <https://doi.org/10.1016/j.arcped.2018.06.008>
- Youngstrom, E. A. (2014). A primer on receiver operating characteristic analysis and diagnostic efficiency statistics for pediatric psychology: We are ready to ROC. *Journal of Pediatric Psychology*, 39(2), 204–221. <https://doi.org/10.1093/jpepsy/jst062>
- Zander, E., Willfors, C., Berggren, S., Choque-Olsson, N., Coco, C., Elmund, A., Moretti, A. H., Holm, A., Jifält, I., Kosieradzki, R., Linder, J., Nordin, V., Olafsdottir, K., Poltrago, L., & Bölte, S. (2016). The objectivity of the autism diagnostic observation schedule (ADOS) in naturalistic clinical settings. *European Child & Adolescent Psychiatry*, 25(7), 769–780. <https://doi.org/10.1007/s00787-015-0793-2>

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