Journal of Human Services: Training, Research, and Practice

Volume 8 | Issue 2 Article 1

10-2022

Autism assessment in the schools: A review of rating scales and observation systems.

Jaime Flowers
Stephen F. Austin University, flowers.jaime.marie@gmail.com

Dallin Marr
Stephen F. Austin State University, dallinmarr@gmail.com

Nina Ellis-Hervey
Stephen F. Austin State University, ellishernm@sfasu.edu

Trudy Georgio
Texas A&M University, georgiobcba@gmail.com

Jessica Cuitareo
cuitareojn@jacks.sfasu.edu

Recommended Citation

Flowers, Jaime; Marr, Dallin; Ellis-Hervey, Nina; Georgio, Trudy; and Cuitareo, Jessica (2022) "Autism assessment in the schools: A review of rating scales and observation systems.," *Journal of Human Services: Training, Research, and Practice*: Vol. 8: Iss. 2, Article 1.
Available at: https://scholarworks.sfasu.edu/jhstrp/vol8/iss2/1

Abstract

School psychologists are tasked with assessing students with Autism Spectrum Disorders (ASD). While not used alone, ASD measures can help practitioners make informed decisions regarding special education eligibility. The purpose of this paper is purpose of the paper is to provide school psychologists and other assessment professionals with a comparison of measures that will aid in selecting the most suitable assessment for a given situation. The following measures were reviewed: Autism Diagnostic Interview, Revised (ADI-R); Autism Diagnostic Observation Schedule, Second Edition (ADOS-2); Autism Spectrum Rating Scale (ASRS); Childhood Autism Rating Scale, Second Edition (CARS-2); and Gilliam Autism Rating Scale, Third Edition (GARS-3).

Keywords: autism assessment, autism, assessment, school assessment, ASRS, ADOS-2, CARS, GARS-3, ADI-R

Autism Assessment in the Schools: A Review of Rating Scales and Observation Systems

According to U.S. National Samples, the prevalence of Autism Spectrum Disorder (ASD) is rising (Liptak et al., 2006). ASD refers to a neurodevelopmental condition associated with challenges in communication, social interactions, and behavioral complications (Thabtah & Peebles, 2019). Currently, about 1 in 44 children has been identified with ASD, according to estimates from the Center for Disease Control's Autism and Developmental Disabilities Monitoring (ADDM) Network. With the growing number of identified cases, there has been some concern surrounding the accuracy, timing, and efficiency of autism diagnosis. To improve the accuracy and reliability of autism diagnoses, experts have developed screening methods to help identify autistic behaviors, speed up the clinical diagnosis referral process, and understand ASD for parents, caregivers, teachers, and family members. However, research studies have demonstrated variability in the screening tools' functionality, accuracy, and reliability, raising questionable complications (Thabtah & Peebles, 2019).

The assessment of students with ASD is a task required of school psychologists. As part of a multidisciplinary team, school psychologists provide expertise in psychopathology and assessment, which is crucial to accurate identification (Aiello et al., 2017). An assessment for ASD typically involves four components: developmental history, interviews, observations, and testing (Esler & Ruble, 2015; Ozonoff et al., 2005). Using appropriate measures with solid reliability and validity is best practice, as valid assessment is essential for informing or verifying diagnosis, evaluating children's strengths and needs, monitoring progress, and developing intervention plans and supports (Paynter, 2015). Failure to diagnose ASD correctly can result in limited resources for students who need the services; a false positive diagnosis can create stress and confusion for the student and their family (Randall et al., 2018).

When surveyed about assessment practices, 92% of school psychologists report being involved in ASD assessment (Aiello et al., 2020). When graduate students are surveyed on training in ASD assessment, only 15% reported that their training in ASD assessment was adequate. Aiello and colleagues (2020) defined evidenced-based assessment of ASD as using a diagnostic measure, intelligence measure, adaptive functioning measure, and a social-emotional/behavioral measure. Less than 25% of school psychologists engaged in evidence-based assessment for ASD. Researchers found most school psychologists are using an ASD rating scale or checklist as the primary tool in assessing ASD.

Understanding the reliability and validity of ASD rating scales is crucial to practitioners assessing the school system. Practitioners heavily rely on these types of measures during ASD assessment.

When looking at which ASD-specific measures are being utilized in practice, Statistics revealed that approximately 82% of the professionals commonly used the following instruments monthly: the CARS-2 (M = 0.56), GARS-3 (M = 0.43), and ADOS-2 (M = 0.43). The most used assessments for ASD are the CARS-2, followed by the ASRS, GARS-3, ADOS-2, ADI-R (Benson et al., 2019). Therefore, evaluating these assessment tools is warranted due to how often these instruments are used.

Autism Diagnostic Interview, Revised (ADI-R) is administered as a semi-structured interview. Other commonly used measures are the Autism Spectrum Rating Scale (ASRS) and Gilliam Autism Rating Scale (GARS-3). These rating scales can also be administered as semi-structured interviews. However, apart from interviews, the Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) is a semi-structured observation that can collect data on ASD-like behavior. Finally, the Childhood Autism Rating Scale, Second Edition (CARS-2), is an evidence-based measure that can be administered as an interview, observation, or combination of the two. None of these measures should be used alone to diagnose ASD but should be used as single elements of a multifaceted assessment process (Esler & Ruble, 2015; Johnson & Myers, 2007; Ozonoff et al., 2005; Wilkinson, 2014). Clinical judgment is essential, especially for differentiating signs of ASD from other disorders with similar presentations (Maddox et al., 2017; Reaven et al., 2008). Another essential concept is the clinical utility of the assessment, which according to the American Psychological Association, is described as "the extent to which diagnostic testing is useful in facilitating beneficial health outcomes from interventions that are initiated based on test results" (VandenBos & American Psychological Association Staff, 2015).

Current Project

Prior to utilizing autism rating scales, professionals must recognize the strengths and weaknesses of the intended scale. The purpose of this paper is to determine the psychometric strengths and weaknesses of the following measures: Autism Diagnostic Interview, Revised (ADI-R; Rutter et al., 2003); Autism Diagnostic Observation Schedule, Second Edition (ADOS-2; Lord et al., 2012); Autism Spectrum Rating Scale (ASRS; Goldstein & Naglieri, 2010); Childhood Autism Rating Scale, Second Edition (CARS-2; Schopler et al., 2010); and Gilliam Autism Rating Scale, Third Edition (GARS-3; Gilliam, 2013). The paper identifies content and use, standardization sample and norms, scores and interpretation, and psychometric properties were reviewed for each measure. A review of each measure will then provide recommendations that may be utilized in practice.

Guidelines for Evaluating ASD scales

Evaluating any rating scale is a multi-factored process. We drew upon several sources (for further discussions, see American Educational Research Association [AERA], et al., 2014;

Bullock & Wilson, 1989; DeVellis, 2016; Edelbrock, 1983; Elliott et al., 1993; McCloskey, 1990) to construct guided criteria from which to judge an instrument's merits. We have condensed the information gleaned into four evaluative dimensions: content and use of a scale, standardization sample and norms, scores and interpretation, and psychometric properties, including reliability and validity. Some of this information is derived from our professional judgment. Two of the authors of this paper are school psychology faculty with 20 years of assessment experience between them.

The content range and scale use include essential aspects for this dimension, such as completeness and user-friendliness of material and manuals, appropriate format (e.g., anchor points, instructions), and scoring procedures. For proper interpretation, norm-referenced measures must be developed with representative standardization samples. Norming procedures should be delineated, including information on the year norming transpired, descriptive statistics, and the sampling procedure used. Scoring and interpretation are the third dimensions. Important aspects of this dimension are detailed descriptions of scores and the appropriateness of scores for the scale. Interpretation of scores should also be delineated and not extend beyond the purposes of the scale (Devillis, 2016).

Psychometric properties look at the reliability and validity of the rating scales. Interrater reliability, test-retest reliability, and internal consistency are essential reliability considerations for most rating scales. Evaluating scale validation includes considerations of content, criterion, and construct validities. For this review, content validity refers to the breadth of diagnostic content covered and how test items were selected. Criterion validity refers to how the test results compare to other measures. Construct validity refers to how well the measure differentiates individuals with ASD from those without.

The criteria of limited, adequate, and excellent are used to characterize data concerning the technical properties of each scale. Limited indicates that a scale is not helpful for research or clinical purposes; adequate indicates a scale may be useful for research or clinical purposes with other data. Excellent indicates the scale is useful for clinical purposes. Because there is no algorithm for determining a given scale or test (AERA, 2014), these criteria and evaluations were guided by the author's judgments of converging evidence for each measure.

Table 1 describes the categories of limited, adequate, and excellent for each dimension we evaluated and indicates where the information comes from. This table was adapted from Hunsley and Mash (2008) .

Procedural Guidelines

Selection criteria for inclusion of the scales were: (a) a specific focus on diagnostic criteria, (b) widespread use in schools, (c) use of the word *autism* in the title, and (d) whether the scale was published at the time of this review. We evaluated widespread use in schools by reviewing the literature on ASD assessment in the school setting. We focused on published scales because we deemed it important to present readily available scales. A previous survey found the GARS, CARS, ADOS, and ADI-R standard measures used in schools (Aiello et al., 2017). The ASRS is a newer nationally normed measure specifically designed for schools. The five scales are presented in alphabetical order. The reviews begin with brief descriptions of each instrument, followed by evaluations based on the four dimensions, focus on diagnostic criteria, widespread use in schools, autism used in the title, and if the scale was published during this review. The authors' also included critical reviews and judgments of the quantity and quality of summary information/data (see Table 2 for descriptive techniques of autism measures).

Autism Diagnostic Interview, Revised Edition

The *Autism Diagnostic Interview, Revised* (ADI-R; Rutter et al., 2003) is semi-structured. The ADI-R is designed for children and adults with a mental age over 2:0. The measure focuses on three domains of functioning: language and communication; reciprocal social interactions; and restricted, repetitive, and stereotyped behaviors and interests. The ADI-R Interview Protocol is composed of an 85-page booklet containing 93 items designed to assist a licensed professional in making a clinical diagnosis. Responses are scored using ADI-R algorithm forms featuring both a diagnostic and a current behavior algorithm. Training is required to administer and code the ADI-R.

The ADI-R remains widely used (Falkmer et al., 2013), though it has not been substantially updated since 1994. Much of the present research on the ADI-R focuses on standardization of the measure in other languages (e.g., de Bildt et al., 2015).

Content and Use

The ADI-R is a semi-structured interview completed by a trained interviewer and an informant (a parent or caregiver knowledgeable about the assessed individual's developmental history and common behavior patterns). The primary focus of the interview is creating a comprehensive developmental account of the client and documenting the current symptom presentation. The interview takes approximately 90 to 150 minutes to administer and score. The interview protocol contains interview questions and coding for up to 42 interview items. The interview items are coded and composited to derive the formal ADI-R algorithm scores. The test additionally features non-algorithmic items, which do not inform diagnosis but provide clinically useful information for intervention development.

Interview questions are organized around the content area, and definitions of all behavioral items are provided. The interview begins with broad introductory background questions followed by questions about the client's development. Next are sections regarding characteristics of the key domains of functioning related to the diagnosis of autism: language and communication functioning, social development and play, and interests and behaviors. The final section contains questions about behaviors of clinical importance.

Standardization Sample and Norms

The manual describes seven reliability and validity studies, with sample sizes ranging from 22 to 94 (total *N* = 335). The ADI-R authors conducted two studies that included 50 and 94. However, reliability statistics were only run on 20 participants in each study. All seven studies used participants from outside of the United States. One study used a German translation of the ADI-R, and another used a Bulgarian translation. Most samples are described in terms of age, disability, and IQ. Demographics for ethnicity are not provided. Only ten female participants are listed in the German or Bulgarian language samples. The samples used the 1994 version of the ADI-R, not the published 2003 version. The ADI-R is a criterion-referenced test, using cutoff algorithm scores instead of norms. The criteria seem to be based on the samples from the authors' two studies.

Scores and Interpretation

The ADI-R is coded using a separate comprehensive algorithm form. Coding is recorded in the corresponding box before the assessor continues to the next interview item. For each item, the clinician gives a code ranging from 0 to 3. A code of 0 is recorded when specified behavior is not reported; a code of 1 is recorded when selected behavior is reported, but not frequent or severe enough to meet the established criteria; a code of 2 indicates abnormal behavior meeting the criteria specified, and a code of 3 is assigned only for extreme severity of the specified behavior. These codes are combined into diagnostic algorithm scores. Classification of autism is determined when the algorithm scores from all three content areas of communication, social interaction, and behavior patterns meet or exceed the specified criteria.

Psychometric Properties

Of the seven reliability and validity studies, three studies examined the interrater reliability of the ADI-R. Combined, the studies reported variable Kappas (.31 - .95) and intraclass correlation coefficients (.52 - .97), with a small sample (*n* = 80). Two studies examined the test-retest reliability of the ADI-R. Combined, the studies reported excellent test-retest reliability (.82 - .97) but again had a small sample size (*n* = 53). Internal consistency is not reported.

Three studies examined the discriminative (construct) validity of the ADI-R. The ADI-R cutoff algorithm scores adequately discriminated Autism and Asperger's from a pervasive developmental disorder, intellectual disability, language impairment, conduct disorder, and typical development (Sensitivity: 80 to 96; Specificity: 92 to 100; *n* = 148). The reciprocal social interaction domain seemed to be the best discriminating. The authors do note the ADI-R has difficulty differentiating nonverbal children with intellectual disability from children with ASD. Content and criterion validity were not reported.

Summary

This measure has limited use. The ADI-R provides a semi-structured interview approach to ASD assessment. The strength of the ADI-R is the large amount of qualitative data it provides relative to symptom presentation and developmental history. The weaknesses of the ADI-R are that it requires training to administer and score, the standardization samples are limited in size and representation, the test takes a significant amount of time to administer, and the test has not been substantially updated since 1994. These difficulties may not make it the best choice in terms of clinical utility.

Autism Diagnostic Observation Schedule, Second Edition

The *Autism Diagnostic Observation Schedule, Second Edition (ADOS-2)* is a semi-structured direct assessment. The measure can be used with individuals aged 12 months to adults. The ADOS-2 can be administered by someone trained and practiced in the assessment, "when using the ADOS-2, examiners need to be sufficiently familiar with the activities and codes so that they can focus their attention on observation of the individual being assessed, rather than on administration details" (Lord, Rutter, et al., 2012, p. 6). The examining psychologists may wish to have other professionals observe the administration of the ADOS-2 and provide input from different disciplines (e.g., occupational therapy or speech and language pathology). In such situations, it is recommended the involved professionals code the ADOS-2 separately then come together to reach a consensus score ("FAQ ADOS-2," n.d.).

Each module can be administered in 40 – 60 minutes and includes multiple activities. The ADOS-2 has five modules. The toddler module (Module T) is for individuals aged 12-30 months and has two scales: social affect and restricted and repetitive behaviors. Module 1 is for individuals 30 months and older who do not consistently use phrase speech. Module 2 is for individuals of any age who use simple phrase speech. Module 3 is for verbally fluent individuals. Modules 1-3 have two scales: social affect and restricted and repetitive behaviors. Module 4 is for verbally fluent adolescents and adults and has four scales: communication, reciprocal social interaction, imagination/creativity, stereotyped behaviors, and restricted interest.

Content and Use

The ADOS-2 has by far the longest manual of the measures included in this review (446 pages; the manual for the GARS-3 is only 53 pages). The length can contribute to the ADOS-2 seeming daunting for new assessors. The ADOS-2 is the only test in the present review which uses manipulatives. Most of the materials are provided in the ADOS-2 kit.

The ADOS-2 is a series of semi-structured activities. An individual's behavior is observed during these activities. Before administration, formal training, preparation, and practice are required to become competent at the ADOS-2. Following administration, the assessor codes the behavior on a variety of items. Coding the behaviors requires the assessor to understand the rating system for each item, as each has a possible combination of eight codings. For example, some items will have codes of 0, 1, 2, and 3, each with their qualitative description, and an additional code 7 indicating different abnormal behaviors. The directions for each module are presented on multiple pages.

Standardization Sample and Norms

Lord, Rutter, et al. (2012) describe three validation samples. The combined samples for Modules 1-3 are large (*N* = 1,467 for Module 1; *N* = 534 for Module 2; *N* = 833 for Module 3). Module T has a moderately sized sample (*N* = 182). Module 4 has a small sample (*N* = 45), all taken from the validation sample of the original ADOS. The samples contain some diversity in ethnic makeup (71-91% White). Participants came from 10 sites throughout the United States and Canada. The samples are predominantly male (57-86%).

Scores and Interpretation

Once the assessment is completed, the assessor must convert the coded ratings into algorithm scores for select items. All ratings, conversions, and scoring are done on the protocol, with directions provided. The selected algorithm scores are added to get the total for each scale. For Modules 1-4, the scale totals are combined and converted to provide the ADOS-2 classification and a comparison score. Three classifications are possible: non-spectrum, autism spectrum, and autism. The comparison score indicates relative severity of ASD symptoms: 8-10 is high, 5-7 is moderate, 3-4 is low, and 1-2 is minimal-to-no evidence of ASD. For Module T, the scale totals are combined and converted to provide a Range of Concern classification. Three classifications are possible: little-to-no concern, mild-to-moderate concern, and moderate-to-severe concern. Module T's classification system "reflects the diagnostic uncertainty that often characterizes clinical observation of young children" (Lord, Luyster, et al., 2012, p. 339).

Psychometric Properties

Lord, Rutter, et al. (2012) provide internal consistency, inter-rater reliability, and test-retest reliability. Internal consistency was measured by the intercorrelation of the algorithm scores with scale and overall totals. These intercorrelations were highly variable (-.06-.88) for all five modules. The test-retest reliability of the overall total is excellent for Modules 1-3 (.83-.87, *n* = 23-27 depending on module). It is also excellent for Module T (.86-.95, *n* = 39). The interrater reliability is strong for Modules 1-3 and Module T (.90-.97, *n* = 14-66 depending on module). Test-retest and interrater reliabilities are not reported for Module 4.

Lord, Rutter, et al. (2012) further provides content and construct validity. Content validity was provided with item discriminations and exploratory factor analyses. Construct validity was determined by calculating sensitivity and specificity. Sensitivity (60-98) and specificity (75-100) were strong for most modules, especially when comparing those with autism against individuals not on the spectrum. Overall, the modules were less accurate in differentiating those with non-autism ASD from individuals not on the autism spectrum. When used with nonverbal children with mental ages below 15 months, Module 1 had poor specificity (19-50). ADOS-2 results were not compared against other measures to establish criterion validity.

Summary

The ADOS-2 is one of the few direct assessment tools available to practitioners to assess students with autism. The strengths of the ADOS-2 are the ability to conduct a semi-structured direct assessment for autism, Modules 1-3 have large validation samples, a multidisciplinary team can observe administration, and adequate to strong reliability and validity. The weaknesses of the ADOS-2 are that it requires training to administer correctly and administration itself requires managing a large amount of material. Module 4 has a small validation sample that has not been updated since 1999 and no reported interrater or test-retest reliability. Due to the issues above, this measure is deemed adequate and should only be given as part of comprehensive assessment practice in clinical work.

Autism Spectrum Rating Scales

The *Autism Spectrum Rating Scales* (Goldstien & Naglieri, 2010) are normative rating scales completed by either a parent or teacher of a child. There are six versions: short-forms for ages 2-5 and 6-18, and parent and teacher versions of long-forms for ages 2-5 and 6-18. The two short forms have 15 items each. The short-forms have only one scale: the short-form scale. The long-form for ages 2-5 has 70 items, with the following twelve scales: ASRS total, social/communication, unusual behaviors, peer socialization, adult socialization, social/emotional reciprocity, atypical language, stereotype, behavioral rigidity, sensory sensitivity, attention/self-regulation, and Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5; the original ASRS had a DSM-IV-TR scale; "ASRS DSM-V scoring update," 2014). The long-form for ages 6-18 has 71 items, with the same scales as the ages 2-5 form, but attention and self-regulation are separated. The different forms allow the ASRS to be used for different purposes. The authors state, "The ASRS helps guide diagnostic decisions and can be used during treatment planning, ongoing monitoring of response to intervention, and program evaluation" (Goldstien & Naglieri, 2010, p. 10).

Content and Use

The ASRS is a rating scale allowing parents and teachers to describe the frequency of a child's behavior related to ASD using a five-point scale, from 0-Never to 4-Very Frequently. Ratings are based on the previous four weeks. Completing the ASRS takes about 20 minutes for the long forms and 5 minutes for the short forms. There are two options for the paper protocols for all versions: the standard form (front and back of one page) or a QuikScore booklet, which facilitates hand scoring. The ASRS can also be administered online. For nonverbal individuals, items related to speech are not rated, and a prorated scoring method is used.

Standardization Sample and Norms

The ASRS has large standardization samples for the long forms (*n* = 640 for ages 2-5; *n* = 1920 for ages 6-18), 50% parent ratings, and 50% teacher ratings. ASRS samples were matched to the U.S. Census data on ethnicity (57.5-62.2% White) and gender (50% male). The samples contain 4.4-8.7% of individuals with ASD. The sample for the short-forms seems to be a subset of the long-form standardization sample (two samples, each with *n* = 695: one for deriving the norms and the second for checking the norms). No specific information about ethnicity, demographics, or gender is provided. However, in describing the construct validity of the short form, a sample of 2,204 is referenced that is 65.7% White.

Scores and Interpretation

Scoring the ASRS can be done right on the protocol (for the QuickScore version), with the scoring software, or online. Interpretation involves examining the scale scores, with the ASRS total and DSM scales being of prime importance. The interpretive framework of the ASRS rests primarily on T-Scores for the various scales. T-Scores below 40 are considered Low, 40 to 59 are Average, 60 to 64 are Slightly Elevated, 65 to 69 are Elevated, and above 70 is Very Elevated in terms of ASD symptom presentation. Norms are chunked by age groups: 2-5, 6-11, and 12-18.

Psychometric Properties

The ASRS manual provides detailed information about the reliability and validity of the measures. The test-retest and inter-rater reliabilities were conducted with appropriately sized samples (*n* = 56-206 divided into parent and teacher forms). Each of the long -forms had strong internal consistency (.70-.97), test-retest reliability (.70-.93), and interrater reliability for the parent forms (.73-.92). The teacher forms for ages 6-18 had lower interrater reliability (.59-.73). Interrater reliability for the teacher form for ages 2-5 is not provided.

The ASRS has strong validity. The reported content validity is based on the DSM-IV-TR diagnostic criteria for Autism Disorder. However, the authors did not mention utilizing experts to confirm that the items are relevant to the DSM-IV-TR. An exploratory factor analysis was also used. For criterion validity, the ASRS is moderately correlated with the GARS-2 (.41-.68; Gilliam, 2006) and the Gilliam Asperger's Disorder Scale (.49-.61; GADS; Gilliam, 2001). The ASRS has a moderate correlation with the original CARS (.06-.66; Schopler et al., 1986). The ASRS short-form and long forms are strongly correlated (.84-.92). As measured by sensitivity

(.90 - .92) and specificity (.89 - .92), the construct validity is also strong when comparing results of ASD, ADHD, other clinical and general population groups.

The ASRS also used EFA to identify factors utilizing their total sample. The analysis revealed that a two-factor solution was appropriate for the parent and teacher/childcare assessment version aged two to five. The first factor was related to social/communication, and the second factor was labeled as unusual behaviors. However, a three-factor solution was most appropriate when utilizing the parent and teacher version for ages six to 18. Based on the analysis, factors were labeled as social communication, unusual behaviors, and self-regulation. Any items excluded from the analysis was based on factor loadings less than .30.

Summary

The ASRS is a rating scale with several options to tailor its use to the rated individual, the rater, and the assessor. It has ages 2-5 and 6-18 forms, long and short forms, parent and teacher forms, and paper and online versions. An alternative scoring method is also available for nonverbal individuals. The strengths of the ASRS are the flexibility previously mentioned, brief administration time, large and diverse normative samples for the long forms, and strong reliability and validity. The weaknesses of the ASRS include the small proportion of the sample which had ASD diagnoses, unclear information about the normative sample for the short forms, and that as a rating scale, the results of the ASRS are biased to the rater's perspective. Being a rating scale also limits the breadth of information provided, especially compared to the qualitative data the ADI-R and ADOS-2 provide.

The ASRS meets our criteria for excellence and is a good choice for ASD assessment.

Childhood Autism Rating Scale-2

The CARS-2 is a 15-item criterion-based and normative rating scale which can assist practitioners in identifying individuals with ASD. The CARS-2 has two versions: CARS-2 Standard Version (CARS2-ST) and CARS-2 High Functioning Version (CARS2-HF). The standard version is used for individuals under the age of 6 or over the age of 6 with an IQ of 70 or lower. The high-functioning version is for individuals over 6 with an IQ over 70. The CARS-2 can be used on individuals aged 3-22.

The CARS-2 can be completed by professionals such as "special educators, school psychologists, speech pathologists, and audiologists, who have had exposure to and training about autism" (Schopler et al., 2010, p. 5). The assessor can incorporate direct observations and interviews of parents and teachers in assigning ratings. A *Questionnaire for Parents or Caregivers* can be given directly to the parents and then used by a professional to complete and score the CARS-2.

Content and Use

The CARS-2 is a rating scale that uses a 4-point rating system, with the option for the assessor to give a .5 if the individual's behavior falls between two points (for a total of 7 response options). Rating options vary from 1 (within normal limits for that age) to 4 (severely abnormal). Each item also has qualitative descriptions for the rating options. There is a space for the assessor to write observation notes for each item. Scoring can be done entirely on the CARS-2 form, and additional information is available in the manual to assist scoring. The CARS-2 can be completed in 10-15 minutes.

Standardization Sample and Norms

The verification samples for the two versions are adequate in terms of size (*n* = 994 – 1034) and ethnicity (60-73% White). The sample has more males than females for both forms. The authors justify this by stating more males than females are diagnosed with autism. The sample is also adequate in the geographic makeup (compared to the 2000 U.S. Census; Schopler et al., 2010). All participants in the sample for the CARS2-ST had previous diagnoses of ASD and IQs below 85. For the CARS2-HF, 58% of participants had previous diagnoses of ASD, and all had IQs above 80.

Scores and Interpretation

Ratings for all fifteen items are totaled, and the total raw score is compared against a criterion and converted into a T-Score and percentile. The scoring can be completed on the first page of the CARS-2 form. Both versions provide an interpretation guide for criterion comparison. On the CARS2-HF, the guide states: 15-27.5 is minimal-to-no symptoms of ASD, 28-33.5 is mild-to-moderate symptoms of ASD, and 34 or higher is severe symptoms of ASD. The CARS2-ST has a similar guide with different totals. A T-Score of 50 represents an average score for someone with ASD, with higher scores representing more severe symptoms.

Psychometric Properties

The authors of the CARS-2 provide internal consistency as a measure of reliability. The internal consistency measured by item-to-total correlations ranged from limited to excellent on both forms (.43-.88 for items; .93-.96 for total score). Excellent inter-rater reliability is reported for the CARS2-HF (.95 for total score). The authors did not provide other types of reliability, instead of providing the original CARS reliability studies, stating that the two versions were similar. The original CARS had excellent test-retest (.77-.90 for periods ranging from three months to two years) and adequate interrater (.84 for total score) reliabilities.

The authors conducted item and factor analyses on both CARS-2 forms for content validity. The EFA resulted in a three-factor solution: Social Communication, Stereotyped Behaviors, Sensory Sensitivities, and Emotional Reactivity. For criterion validity, results of the CARS-2 were compared against the original ADOS (Lord et al., 1999) and the Social Responsiveness Scale (SRS; Constantino & Gruber, 2005). CARS-2 and ADOS results were strongly correlated (.77-.79). CARS-2 and SRS results had low to moderate correlation (.38-.47). For construct validity, the CARS2-HF is reported to have moderate sensitivity (.81) and specificity (.87), though these values drop when compared to a non-ASD clinical population (sensitivity: .77; specificity: .58). The authors provided evidence of the original CARS sensitivity (.88) and specificity (.86) but not the CARS2-ST.

Summary

The CARS-2 is an adequate measure and should only be used as part of a comprehensive assessment procedure. The CARS-2 combines information from different sources into a 15-item rating scale completed by a professional. The strengths of the CARS-2 are in its incorporation of other sources of information such as direct observation and interviews, brief administration time, and a simple and accessible scoring system. The weaknesses of the CARS-2 are the low sensitivity and specificity compared to clinical populations and the lack of updated reliability and validity information. More reliability and validity studies with the new version are needed to determine the technical adequacy of the updated measure. The CARS-2 is also the only assessment in the current review which has not been translated into Spanish.

Gilliam Autism Rating Scale, Third Edition

The Gilliam Autism Rating Scale, Third Edition (GARS-3; Gilliam, 2014) is a 58-item rating scale used to screen individuals of ages 3-22 for ASD. Responses are made on a 4-point scale. The measure has six subscales: restricted/repetitive behaviors, social interaction, social communication, emotional responses, cognitive style, and maladaptive speech. These subscale scores combine into the Autism Index. An alternate four-scale index can be calculated for individuals who are nonverbal. One specific goal of the developer was to make the measure useable in schools – by parents, teachers, and assessors. The GARS-3 manual states it is based on the changes to the DSM-5 definition of ASD.

Content and Use

Parents or teachers can complete the GARS-3 in about 5-10 minutes. They rate the individual's typical behavior using a four-point scale, from 0-Not at all like the individual to 3-Very much like the individual. Raters should have "regular, sustained contact with the individual for at least two weeks" (Gilliam, 2014, p. 8) and are encouraged if unsure on an item to observe for six hours. The separate *GARS-3 Instructional Objectives* manual provides possible goals, objectives, and interventions based on subscale responses.

Standardization Sample and Norms

The GARS-3 has a large standardization sample (*n* = 1,859), all of whom had previous diagnoses of ASD. The sample's ethnic makeup resembles school-age children reported in the 2010 U.S. census (80% White; Gilliam, 2014). The raters for this sample included teachers, parents of children with ASD, speech clinicians, teacher assistants, psychologists, and educational diagnosticians. Most of the raters self-reported advanced degree attainment (58.6%), high levels of knowledge about ASD (71.4%), and more than six years of experience with individuals with ASD (58.9%). Separate age and gender norms were not provided because the authors found weak correlations.

Scores and Interpretation

Sums of item responses provide the raw scores for each subscale. The manual converts raw scores into subscale scaled scores (M = 10, SD = 3), index standard scores (M = 100, SD = 15), and percentiles. The GARS-3 is hand-scored using the tables in the manual. The protocols have an interpretation guide on the front page and an ASD diagnostic validation checklist on the back page. The ASD diagnostic validation checklist covers DSM-5 criteria A through D. The Autism Index has the following interpretation: standard scores of 71 or higher indicate the probability of ASD as very likely, 55 to 70 indicate the possibility of ASD as probable, and 54 or lower indicate the probability of ASD as unlikely. Using this method, up to three standard deviations below the mean can be considered "probable" for a diagnosis of ASD.

Psychometric Properties

Gilliam (2014) provided internal consistency, test-retest reliability, and interrater reliability. The internal consistency for the overall measure was excellent (.93-.94). The test-retest was conducted with 122 participants one to two weeks apart; results were excellent (.90). Interrater reliability was also excellent (.84 for 116 pairs of raters).

Gilliam (2014) provided content, criterion, and construct validity. For content validity, expert opinion, factor analysis, and item discriminations were used to select the most appropriate items for the GARS-3. For criterion validity, results from the GARS-3 were compared to results from the Autism Behavior Checklist (part of the Autism Screening Instrument for Educational Planning, Third Edition; Krug et al., 2008), the CARS-2 (Schopler et al., 2010), the GADS (Gilliam, 2001), and the ADOS (Lord et al., 1999). Correlations were excellent (.69-.83). For construct validity, participants with autism were found to have higher average GARS-3 autism index scores than most non-ASD peers. Individuals with Intellectual Disability (ID) scored high on the GARS-3 (M = 87-89, *n* = 15), suggesting that the GARS-3 is not reliably able to differentiate between ASD and ID. Sensitivity (.83-.98) and specificity (.62-.97) are moderate when comparing the results of those with ASD against those with other non-ID disabilities and typically functioning individuals.

The GARS used exploratory factor analysis (EFA) to identify the factors within the assessments. Utilizing EFA, the GARS used the entire autism sample collected (*N* = 1,859) to identify factors. When unrotated, EFA revealed a single factor that accounted for 46% of the variance, while the rotated factor solution revealed six factors. The GARS subscales include restricted/repetitive behaviors, social interaction, social communication, emotional responses, cognitive style, and maladaptive speech.

Summary

The GARS-3 is a rating scale that serves as a screener for ASD. The GARS-3 meets the criteria for excellent and is another good choice as an assessment tool. The strengths of the GARS-3 are the brief administration time, large and diverse standardization sample, and strong reliability and validity. The weaknesses of the GARS-3 are a poor ability to differentiate between ASD and similar disorders like ID, probably related to the overly generous cutoff scores of the GARS-3, and the normative sample consisting of a majority of highly educated raters with a lot of experience with ASD. It also suffers the same rating scale drawbacks as the ASRS: susceptibility to the rater's bias and lack in the breadth of qualitative information.

Conclusions

Practitioners using autism rating scales need to identify the strengths and weaknesses of the instrument they choose to use. Thus, potential autism rating scale users need to be informed and skeptical consumers. The following recommendations are offered by evaluating the areas on which we chose to focus (content and use, standardization sample and norms, scores and interpretation, and reliability and validity).

The most comprehensive instrument is the ASRS. We recommend using it for a strong norming population and strong reliability and validity. We recommend using the GARS-3 or the ASRS short form for a shorter screener. Both have adequate psychometric properties and can be administered in five minutes. To include a direct structured measure in an assessment, we recommend using the CARS-2, which can be used as an observation guideline. The CARS-2 has a strong norming population. However, caution is required because most of the evidence of reliability and validity is provided through the original version of the CARS and not the newer version. CARS-2 should only be used in conjunction with other ASD assessment tools.

The ADOS-2 is not recommended due to how difficult it is to administer. A direct standardized assessment that provides activities with the child is a useful tool; however, the ADOS-2 is simply difficult to use and learn. If a practitioner has been trained in ADOS-2 administration the ADOS-2 may be a useful tool for ASD assessment, however should only be used in conjunction with other measures due to its limitations. ADI-R is not recommended for use due to its limited reliability and validity.

In conclusion, many important aspects must be considered when evaluating the quality and usefulness of rating scales. Users must be knowledgeable about the instrument's qualities for appropriate use and interpretation. Researchers should focus on evaluating commonly used assessment tools and providing recommendations to practitioners. Practitioners should become familiar with the evaluating assessment tools, so that they can ensure they are using evidence-based assessment.

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

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for educational and psychological testing*. AERA.

Autism spectrum rating scales (ASRS) – DSM-V scoring update. (2014, November 18). Pearson. https://www.pearsonclinical.co.uk/Psychology/ChildMentalHealth/ChildAutisticSpectrumDisorders/AutismSpectrumRatingScales/ForThisProduct/asrs-dsm-v-update.aspx

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

Bullock, L. M., & Wilson, M. J. (1989). *Behavior dimensions rating scale: Examiner's manual*. DLM Teaching Resources.

Centers for Disease Control and Prevention. (2021, December 2). *Data & statistics on autism spectrum disorder*. Centers for Disease Control and Prevention. Retrieved January 1, 2022, from https://www.cdc.gov/ncbddd/autism/data.html

Constantino, J. N., & Gruber, C. P. (2005). *Social responsiveness scale (SRS)*. Western Psychological Services.

De Bildt, A., Sytema, S., Zander, E., Bölte, S., Sturm, H., Yirmiya, N., Yaari, M., Charman, T., Salomone, E., LeCouteur, A., Green, J., Bedia, E. C., Primo, P. G., van Daalen, E., de Jonge, M. V., Guomundsdottir, E., Johannsdottir, S., Raleva, M., Boskovska, M., … & Oosterling, I. J. (2015). Autism diagnostic interview – revised (ADI-R) algorithms for toddlers and young preschoolers: Application in a non-US sample of 1,104 children. *Journal of Autism and Developmental Disorders*, *45*, 2076-2091. https://10.1007/s10803-015-2372-2

DeVellis, R. F. (2016). *Scale development: Theory and applications* (4th ed.). Sage Publications, Inc.

Edelbrock, C. (1983). Problems and issues in using rating scales to assess child personality and psychopathology. *School Psychology Review, 12*, 293-299. https://doi.org/10.1080/02796015.1983.12085045

Elliott, S. N., Busse, R. T., & Gresham, E M. (1993). Behavior rating scales: Issues of use and development. *School Psychology Review, 22*, 313-321. https://doi.org/10.1080/02796015.1993.12085655

Esler, A. N., & Ruble, L. A. (2015). DSM-5 diagnostic criteria for autism spectrum disorder with implications for school psychologists. *International Journal of School & Educational Psychology, 3*(1), 1–15. https://doi.org/10.1080/21683603.2014.890148

Falkmer, T., Anderson, K., Falkmer, M., & Horlin, C. (2013). Diagnostic procedures in autism spectrum disorders: A systematic literature review. *European Child and Adolescent Psychiatry, 22*, 329–340. https://doi.org/10.1007/s00787-013-0375-0

FAQ ADOS-2 team-based autism assessment. (n.d.). Western Psychological Services. https://www.wpspublish.com/faq-ados-2-team-based-autism-assessment

Gilliam, J. E. (2001). *Gilliam Asperger's disorder scale*. Pro-Ed.

Gilliam, J. E. (2006). *Gilliam autism rating scale* (2nd ed.). Pro-Ed.

Gilliam, J. E. (2014). *Gilliam autism rating scale* (3rd ed.). Pro-Ed.

Goldstein, S., & Naglieri, J. A. (2010). *Autism spectrum rating scales*. Multi-Health Systems Inc.

Hunsley, J., & Mash, E. J. (Eds.). (2008). A guide to assessments that work (1st ed.). Oxford University Press.

Johnson, C. P., & Myers, S. M. (2007). Identification and evaluation of children with autism spectrum disorders. *Pediatrics*, *120*(5), 1183. https://doi.org/10.1542/peds.2007-2361

Krug, D. A., Arick, J. R., & Almond, P. J. (2008). *Autism screening instrument for educational planning* (3rd ed.). Pro-Ed.

Lord, C., Luyster, R. J., Gotham, K., & Guthrie, W. (2012). *Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) Manual (Part II): Toddler Module.* Western Psychological Services.

Lord, C., Rutter, M., DiLavore, P., & Risi, S. (1999). *Autism diagnostic observation scale*. Psychological Services.

Lord, C., Rutter, M., DiLavore, P., Risi, S., Gotham, K., & Bishop, S. L. (2012). *Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) Manual (Part I): Modules 1-4.* Western Psychological Services.

MacFarlane, J. R., & Kanaya, T. (2009). What does it mean to be autistic? Inter-state variation in special education criteria for autism services. *Journal of Child and Family Studies, 18,* 662-669. https://doi.org/10.1007/s10826-009-9268-8

Maddox, B. B., Brodkin, E. S., Calkins, M. E., Shea, K., Mullan, K., Hostager, J., & Miller, J. S. (2017). The accuracy of the ADOS-2 in identifying Autism among adults with complex psychiatric conditions. *Journal of Autism and Developmental Disorder*, *47(9)*, 2703-2709. https://doi.org/10.1007/s10803-017-3188-z

McCloskey, G. (1990). Selecting and using early childhood rating scales. *Topics in Early Childhood Special Education, 10(3)*, 39-64. https://doi.org/10.1177/027112149001000305

Molloy, C. A., Murray, D. S., Akers, R., Mitchell, T., & Manning-Courtney, P. (2011). Use of the Autism Diagnostic Observation Schedule (ADOS) in a clinical setting. *Autism*, *15*(2), 143-162. https://doi.org/10.1177/1362361310379241

Ozonoff, S., Goodlin-Jones, B. L., & Solomon, M. (2005) Evidence-based assessment of autism spectrum disorders in children and adolescents. *Journal of Clinical Child and Adolescent Psychology*, *34*(3), 523-540. https://doi.org/10.1207/s15374424jccp3403_8

Paynter, J. (2015). Assessment of School-Aged Children with Autism Spectrum Disorder. *Journal of Psychologists and Counsellors in Schools, 25*(1), 104-115. https://doi.org/10.1017/jgc.2015.2

Randall, M., Egberts, K. J., Samtani, A., Scholten, R. J., Hooft, L., Livingstone, N., Sterling-Levis, K., Woolfenden, S., & Williams, K. (2018). Diagnostic tests for autism spectrum disorder (ASD) in preschool children. Cochrane Database of Systematic Reviews, 7, Article CD009044. https://doi.org/10.1002/14651858.CD009044.pub2

Reaven, J. A., Hepburn, S. L., & Ross, R. G. (2008). Use of the ADOS and ADI-R in children with psychosis: Importance of clinical judgment. *Clinical Child Psychology and Psychiatry*, *13*(1), 81–94. https://doi.org/10.1177/1359104507086343

Rutter, M., LeCouteur, A., & Lord, C. (2003). *Autism diagnostic interview - revised*. Western Psychological Services.

Schopler, E., Reichler, R. J., & Rochen Renner, B. (1986). *Childhood autism rating scale*. Western Psychological Services.

Schopler, E., Van Bourgondien, M. E., Wellman, G. J., & Love, S. R. (2010). *Childhood autism rating scale* (2nd ed.). Western Psychological Services.

Thabtah, F., & Peebles, D. (2019). Early Autism Screening: A Comprehensive Review. *International journal of environmental research and public health*, *16*(18), 3502. https://doi.org/10.3390/ijerph16183502

VandenBos, G. R., & American Psychological Association Staff. (2015). *APA concise dictionary of psychology*. American Psychological Association.

Volker, M. A., & Lopata, C. (2008). Autism: A review of biological bases, assessments, and intervention. *School Psychology Quarterly*, *23*(2), 258-270. https://10.1037/1045-3830.23.2.258

Wilkinson, L. A. (Ed.). (2014). *Autism spectrum disorder in children and adolescents: Evidence-based assessment and intervention in schools*. American Psychological Association.

Table 1: *Guidelines used to evaluate measures*

	Limited	Adequate	Excellent
Content and use *(Professional judgement of authors)*	Unclear or missing information	Has all information, but may not be presented in a clear use friendly style	Complete User-Friendly
Standardization sample and norms *(Professional judgement of authors; AERA, 2014)*	Information is missing or sample is not small or representative of the population of intended use	Sample is either large or representative of the population of intended use	Sample is both large and representative of the population of intended use
Scores and interpretation *(Professional judgement of authors)*	Scoring and interpretation guidelines are not clear	Scoring or interpretation guidelines are fairly clear but could use some more description	Reliability and validity information Clear scoring and interpretation guidelines are very clear
Reliability *(Devillis, 2016)*	<0.60	0.61-0.89	>0.90
Validity *(DeVellis, 2016; Edelbrock, 1983; Elliott et al., 1993)*	2-week test-retest < 0.60 or not described. Other forms of validity not described or unclear	2-week test-retest 0.61-0.79 Utilized exploratory or confirmatory factor analysis Criterion or construct validly described	2-week test-retest ≥ .80 Utilized both exploratory and confirmatory factor analysis. Criterion and construct validly described

Table 2:
Descriptive Techniques of Autism Measures

Test Name	Items	Age Range and Normative Sample	Interpretive Profile	Response Format
Autism Diagnostic Interview - Revised (ADI-R; Rutter, LeCouteur, and Lord, 2003)	93	Children and adults with a mental age above 2:0 Reliability data comes from the 1994 version of the ADI-R Validation Samples (*N* = 338) ~50% with Autism No ethnic demographics provided Only 5 female participants reported; unclear on gender breakdown of whole sample	3 domains Language/Communication Reciprocal Social Interactions Repetitive Behaviors/Interests Determines diagnostic suggestion with an algorithm based on cutoff scores in each domain	Semi-structured interview of parent or caregiver 4-point Likert scale 0 is normal 3 is very abnormal
Autism Diagnostic Observation System, Second Edition: Toddler Module (ADOS-2; Lord, Luyster, et al., 2012)	11	12-30 months Validation Sample (*N* = 182) 25% with ASD 80% Caucasian 76% Male	2 scales Social Affect Restricted and Repetitive Behaviors	Semi-structured observational assessment of individuals suspected of having ASD 3 to 4-point Likert scale, varies based on items
Autism Diagnostic Observation System, Second Edition: Modules 1-4 (ADOS-2; Lord, Rutter, et al., 2012)	10-15	31 months to adult Validation sample included original ADOS sample Validation Samples (*N* = 3,293)	Modules 1-3: 2 scales Social Affect Restricted and Repetitive Behaviors Module 4: 4 scales Communication	Same as ADOS-2 Toddler Module

		32-91% with Autism 0-47% Non-Autism ASD 71-91% Caucasian 57-86% Male	Reciprocal Social Interaction Imagination/Creativity Stereotyped Behaviors and Restricted Interest	
Autism Spectrum Rating Scales (ASRS) (2-5 years) (Goldstein & Naglieri, 2010)	70	2-5 years Normative Sample (*N* = 320) 4.4% with Autism. 62.2% Caucasian 50% Male	2 ASRS scales Social/Communication Unusual Behaviors 8 treatment scales Peer Socialization Adult Socialization Social/Emotional Reciprocity Atypical Language Stereotype Behavioral Rigidity Sensory Sensitivity Attention/Self-Regulation 1 DSM Scale	Rating scales for teachers/childcare providers and parents. 5-point Likert scale 0 is never 1 is rarely 2 is occasionally 3 is frequently 4 is very frequently
Autism Spectrum Rating Scales (ASRS) (6-18 year) (Goldstein & Naglieri, 2010)	71	6-18 years Normative Sample (*N* = 960) 8.7% with Autism 68.2% Caucasian 50% Male	3 ASRS scales Social/Communication Unusual Behaviors Self-Regulation 8 treatment scales Peer Socialization Adult Socialization Social/Emotional Reciprocity Stereotype Behavioral Rigidity Sensory Sensitivity Attention 1 DSM Scale	Same as ASRS full length (2-5 years)

Child Autism Rating Scale, Second Edition – Standard (CARS-2 ST; Schopler, Van Bourgondien, Wellman & Love, 2010)	15	0-6 or over 6 with IQ of 79 or lower. Validation Sample (*N* = 1,034) 100% with Autism and IQ below 85 60% Caucasian 78% Male 30% Under 6	1 scale Severity Group	Rating scale completed by well-informed professional based on interview and/or observation data 4-point Likert scale varies based on items, with ability to rate the item .5 if the person's abilities fall between two points
Childhood Autism Rating Scale, Second Edition – High Functioning (CARS-2 HF; Schopler et al., 2010)	15	6 or older with IQ of 80 or higher Validation Sample (*N* = 994) 58% with ASD 73% Caucasian 78% Male	1 scale Severity Group	Same as CARS-2 ST.
Gilliam Autism Rating Scale, Third Edition (GARS-3; Gilliam, 2013)	58	3-22 years Normative Sample (*N* = 1,859) 100% diagnosed with ASD 80% Caucasian 77% Male	6 Subscales Restricted/Repetitive Behaviors Social Interaction Social Communication Emotional Responses Cognitive Style Maladaptive Speech Autism Index 4 Autism Index 6	Rating scale completed by teacher, parent, or other caregiver who has had more than 2 weeks contact with individual suspected as having ASD 4-point Likert scale 0 Not at all like the individual 1 Not much like the individual 2 Somewhat like the individual 3 Very much like the individual

Table 2:
Reliability of Autism Measures

Test Name	Internal Consistency	Test-Retest	Interrater
ADI-R (Rutter et al., 2003)	None provided	*n* = 20 Scales: .93-.97 (2-5 months) *n* = 33 Scales: .82-.91 (4-6 weeks)	*n* = 42 Items: .37-.95 *n* = 38 Scales: .59-.87
ADOS-2 Toddler Module (Lord, Luyster, et al., 2012)	Intercorrelations of algorithm items with domain and overall totals: Younger or few words: .11-.85 Older with some words: .18-.81	.86-.95 (length of time not provided)	.90-.99
ADOS-2 Modules 1-4 (Lord, Rutter, et al., 2012)	Intercorrelations of algorithm items with domain and overall totals: Module 1 (≤15 months): -.06-.65 Module 1 (>15 months): .20-.74 Module 1 (some words): .38-.78 Module 2 (<5 years): .24-.71 Module 2 (≥5 years): .28-.77 Module 3: .08-.72 Module 4: .23-.88	Module 1-3: 83-.87 (average of 10 months; *n* = 23-27) Module 4: none provided	Module 1-3: .94-.97 overall (*n* = 50-66) Module 4: none provided
ASRS (2-5 years) (Goldstein & Naglieri 2010)	Parent Ratings: .74-.97 Teacher Ratings: .70-.97	(*N* = 56) Parent Ratings: .79-.93 (2-4 wks.) (*N* = 62) Teacher Ratings: .72-.92 (2-4 wks.)	(*N* = 64) Parent Ratings: .73-.87 Teacher Ratings: not provided
ASRS (6-18 year) (Goldstein & Naglieri, 2010)	6-12 years Parent Ratings: .77-.97 Teacher Ratings: .73-.97 12-18 years	(*N* = 109) Parent Ratings: .87-.92 (2-4 wks.) (*N* = 218) Teacher Ratings: .70-.88 (2-4 wks.)	(*N* = 84) Parent Ratings: .83-.92 (*N* = 115) Teacher Ratings: .59-.73

	Parent Ratings: .78-.97 Teacher Ratings: .77-.98		
CARS-2 ST (Schopler et al., 2010)	Item-to-total correlations: .43-.81 Total: .93	Not provided	Not Provided
CARS-2 HF (Schopler et al., 2010)	Item-to-total correlations: .53-.88 Total: .96	Not Provided	*N* = 239 .51-.90 Total: .95
GARS-3 (Gilliam, 2013)	Subscales (by age): .71-.96 Autism Index 4: .94 Autism Index 6: .93	*n* = 122 (one to two weeks apart) Subscales: .76-.87 Autism Index 4: .90 Autism Index 6: .90	232 raters (116 pairs) Subscales: .71-.85 Autism Index 4: .84 Autism Index 6: .84

Table 3

Validity of Autism Measures

Test Name	Content	Criterion Related	Construct
ADI-R (Rutter et al., 2003)	The ADI-R includes items relevant to domains A and B from the DSM-5	None provided	3 of the validation studies found the ADI-R to differentiate well between Autism and other disabilities. The studies also found good specificity and sensitivity
ADOS-2 Toddler Module (Lord, Luyster, et al., 2012)	Item Discrimination produced strong sensitivities and specificities in all areas	None provided	77-84% of individuals with autism were scored as moderate to severe concern 82-92% of individuals who were non spectrum or typically developing were scored as little to no concern
ADOS-2 Modules 1-4 (Lord, Rutter, et al., 2012)	Item Discrimination produced strong sensitivities and specificities in all areas Factor analysis was conducted on the ADOS-2 to confirm the subscale structure	None provided	Modules 1-3 had differential scoring between disability groups. The differences between groups in Module 4 were small
ASRS (2-5 years) (Goldstein & Naglieri 2010)	Items are conceptually consistent with key symptomatic areas of autism spectrum disorder according to multiple sources	*Gilliam Rating Scale* (Gilliam, 1995) Parent rating (*N* = 78): .61 Teacher rating (*N* = 53): .41 *Childhood Autism Rating Scale*	Teacher and parent ratings demonstrated differential scoring between clinical groups

	The ASRS scales of the full-length form were developed through an exploratory factor analysis	(Schopler et al., 1998) Parent rating (*N* = 34): .66 Teacher rating (*N* = 36): .06 *Gillian Asperger's Disorder Scale* (Gilliam, 2001) Parent rating (*N* = 78): .49 Teacher rating (*N* = 52): .56	
ASRS (6-18 years) (Goldstein & Naglieri, 2010)	Same as ASRS (2-5 years)	*Gilliam Rating Scale* (Gilliam, 1995) Parent rating (*N* = 104): .63 Teacher rating (*N* = 116): .68 *Childhood Autism Rating Scale* (Schopler et al., 1998) Parent rating (*N* = 109): .54 Teacher rating (*N* = 122): .61 *Gillian Asperger's Disorder Scale* (Gilliam, 2001) Parent rating (*N* = 83): .40 Teacher rating (*N* = 82): .51	Same as ASRS (2-5 years)
CARS-2 ST (Schopler et al., 2010)	Item Discriminations were used to eliminate items which were not useful Factor analysis was conducted on the CARS-2	ADOS (Lord et al., 1999) *N* = 37, r = .79 Social Responsiveness Scale (SRS; Constantino & Gruber, 2005) *N* = ?, r = .38	None provided

CARS-2 HF (Schopler et al., 2010)	Same as CARS-2 ST	ADOS (Lord et al., 1999) *N* = 76, r = .77 Social Responsiveness Scale (SRS; Constantino & Gruber, 2005) *N* = 293, r = .47	465 of the 520 individuals who scored in the high range (28 or more) had a clinical diagnosis of ASD
GARS-3 (Gilliam, 2013)	Items were based on the DSM-5, other autism measures, and expert opinion. Factor analysis, item discrimination, and item analysis were also used	Autism Behavior Checklist (Krug, Arick & Almond, 2008) *N* =74, r = .76-.86 CARS-2 (Schopler et al., 2010) *N* =128, r = .66-.83 Gilliam Asperger's Disorder Scale (Gilliam, 2001) *n* = 61, r = .70-.75 ADOS (Lord et al., 1999) *n* = 56, r = .61-.72	Mean Autism Index 6 of groups with disabilities: ASD: 100 ID (*n* = 15): 87 ADHD (*n* = 73): 55 ED/BD (*n* = 58): 60 LD (*n* = 163): 51 SLI (*n* = 54): 59 No disability (*n* = 130): 50 Sensitivity: .83-.98 Specificity: .62-.97