CHAPTER TWO

Psychometric Issues and Current Scales for Assessing Autism Spectrum Disorder

Jack A. Naglieri
Kimberly M. Chambers
Keith D. McGoldrick
Sam Goldstein

The study of any psychological disorder is dependent upon the tools that are used, as these tools directly influence what is learned about the subject in research as well as clinical practice. As in all areas of science, what we discover depends upon the quality of the instruments we use and the information they provide. Better-made instruments yield more accurate and reliable information. Instruments that uncover more information relevant to the subject being examined will have better validity, and ultimately will more completely inform both researchers and clinicians. The tools we use for diagnosis have a substantial impact on the reliability and validity of the information we obtain and the decisions we make. Simply put, the better the tool, the more valid and reliable the decisions, the more useful the information obtained, and the better the services that are eventually provided. In this chapter, the tools used for assessing the characteristics of children and adolescents who have autism spectrum disorder (ASD) are examined.

This chapter has two goals. First, we review the important psychometric qualities of test reliability and validity. The aim of this first section is to illustrate the relevance of reliability and validity for the decisions made by clinicians and researchers whose goal is to better understand ASD. We emphasize the practical implications these psychometric issues have for the assessment of ASD, and the implications they have for interpretation of results within and across instruments. Special attention is also paid to scale development procedures, particularly methods used to

develop derived scores. The second section of this chapter focuses on the various measures used to assess ASD. The structure, reliability, and validity of each instrument are summarized. The overall aim of the chapter is to provide an examination of the relevant psychometric issues and the extent to which researchers and clinicians can have confidence in the tools they use to assess ASD.

PSYCHOMETRIC ISSUES

Reliability

The reliability of any variable, test, or scale is critical for clinical practice as well as research purposes. It is important to know the reliability of a test, so that the amount of accuracy in a score can be determined and used to calculate the amount of error in the measurement of the construct. The higher the reliability, the smaller the error, and the smaller the range of scores that are used to build the confidence interval around the estimated true score. The smaller the range, the more precision and confidence practitioners can have in their interpretation of the results.

Bracken (1987) provided levels for acceptable test reliability. He stated that individual scales from a test (e.g., a subtest or subscale) should have a reliability of .80 or greater, and that total tests should have an internal consistency of .90 or greater. The reason for testing and the importance of the decisions made could also influence the level of precision required—that is, if a score is used for screening purposes (where overidentification is preferred to underidentification), a .80 reliability standard for a total score may be acceptable. However, if decisions are made, for example, about special educational placement, then a higher reliability (e.g., .95) would be more appropriate (Nunnally & Bernstein, 1994).

Every score obtained from any test is composed of the true score plus error (Crocker & Algina, 1986). We can never obtain the true score, so we describe it on the basis of a range of values within which the person's score falls at a specific level of certainty (e.g., 90% probability). The range of scores (called the confidence interval) is computed by first obtaining the standard error of measurement (*SEM*) from the reliability coefficient and the standard deviation (*SD*) of the score in the following formula (Crocker & Algina, 1986):

$$SEM = SD \times \sqrt{1 - \text{reliability}}$$

The confidence interval should be used in practice, to better describe the range of scores that is likely to contain the true score. In practice, we say that a child earned an IQ score of 105 (±5), and state that there is a 90%

likelihood that the child's true IQ score falls within the range of 100–110 (105 ± 5).

The confidence interval is based on the *SEM,* which is the average *SD* of a person's scores around the true score. For this reason, we can say that there is a 68% chance (the percentage of scores contained within ±1 *SD*) that the person's true score is within that range. Recall that 68% of cases in a normal distribution fall within +1 and –1 *SD*. The *SEM* is multiplied by a *z* value of, for example, 1.64 or 1.96, to obtain a confidence interval at the 90 or 95% level, respectively. The resulting value is added to and subtracted from the obtained score to yield the confidence interval. So in the example provided above, the confidence interval for an obtained score of 100 is between 95 (100 – 5) and 105 (100 + 5). Figure 2.1 provides confidence intervals (95% level of confidence) for a standard score of 100 that would be obtained for measures with reliability of .50–.99. As would be expected, the range within which the true score is expected to fall varies considerably as a function of the reliability coefficient, and the lower the reliability, the wider the range of scores that can be expected to include the true score.

Technically, however, the confidence interval (and *SEM*) is centered on the estimated true score rather than the obtained score (Nunnally & Bernstein, 1994). In many published tests—for example, the Wechsler Intelligence Scale for Children, Fifth Edition (Wechsler, 2014) and the Cognitive Assessment System, Second Edition (Naglieri, Das, & Goldstein, 2014)—the confidence intervals are provided in the test manual's table for converting sums of subtest scores to standard scores, and the range is already centered on the estimated true score. The relationships among the various scores are illustrated in Table 2.1, which provides the obtained score, estimated true score, and lower and upper ranges of the confidence intervals

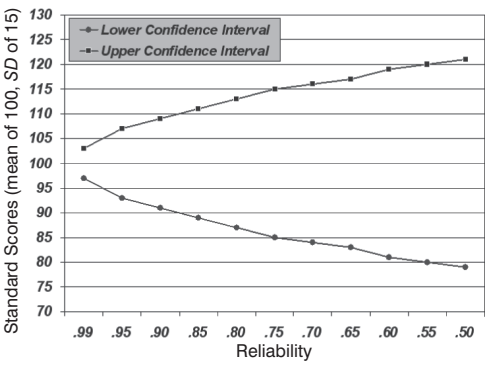

FIGURE 2.1. Relationships between reliability and confidence intervals.

for standard scores (mean of 100, *SD* of 15) for a hypothetical test with a reliability of .90 at the 90% level of confidence.

Examination of these scores shows that the confidence interval is equally distributed around a score of 100 (92 and 108 are both 8 points from the obtained score), but the interval becomes less symmetrical as the obtained score deviates from the mean. For example, ranges for standard scores that are below the mean are *higher* than the obtained score. As shown in Table 2.1, the range for a standard score of 80 is 74–90 (6 points below 80 and 10 points above 80). In contrast, ranges for standard scores that are above the mean are *lower* than the obtained score. The range for a standard score of 120 is 110–126 (10 points below 120 and 6 points above 120). This difference is the result of centering the range of scores on the estimated true score rather than the obtained score. Note that the size of the confidence interval is constant (±8 points) in all instances. Regardless of how the confidence intervals are constructed, the important point is that measurement error must be known and taken into consideration when scores from any measuring system are used. Confidence intervals, especially those that are based on the estimated true score, should be provided for all test scores including rating scales.

The importance of the *SEM* becomes most relevant when two scores are compared. The lower the reliability, the larger the *SEM*, and the more

TABLE 2.1. Relationships among Obtained Standard Scores, Estimated True Scores, and Confidence Intervals across the 40–160 Range

Obtained standard score	Estimated true score	True minus obtained score	Lower confidence interval	Upper confidence interval	Upper minus lower confidence interval
40	46	6	38	54	16
50	55	5	47	63	16
60	64	4	56	72	16
70	73	3	65	81	16
80	82	2	74	90	16
90	91	1	83	99	16
100	100	0	92	108	16
110	109	–1	101	117	16
120	118	–2	110	126	16
130	127	–3	119	135	16
140	136	–4	128	144	16
150	145	–5	137	153	16
160	154	–6	146	162	16

Note. This table assumes a reliability coefficient of .90 and a 90% confidence interval.

likely an individual's scores are to differ on the basis of chance. For example, when a child's score on a measure of selfregulation is compared with scores on a measure of social skills, the reliability of these measures will influence their consistency and therefore the size of the difference between them. The lower the reliability, the more likely they are to be different by chance alone. The formula for determining how different two scores need to be includes the *SEM* of each score and the *z* score associated with a specified level of significance. The difference can be computed by using the following formula:

$$Difference = Z \times \sqrt{SEM1_2 + SEM2_2}$$

The difference needed for significance when one is comparing two variables with reliability coefficients of .85 and .78, using an *SD* of 15, is easily calculated with the formula above. To illustrate, scores on measures of selfregulation (with a reliability of .85) and social skills (reliability of .78) would have to differ by 19 points (more than an entire *SD*) to be significant. Figure 2.2 provides the values that would be needed for comparing two scores with the same reliability, ranging from very good (.95) to very poor (.40) at the .05 level of significance, and a standard score that has an *SD* of 15. This figure shows that when one is comparing two scores with reliabilities of .70, differences of more than 20 points would be attributed to *measurement error alone*. Clearly, in both research and clinical settings, variables with high reliability are needed.

It is therefore important that researchers and clinicians who assess behaviors associated with ASD use measures that have a reliability coefficient of .80 or higher and composite score reliabilities of at least .90. If a test or rating scale does not meet these requirements, then its inclusion

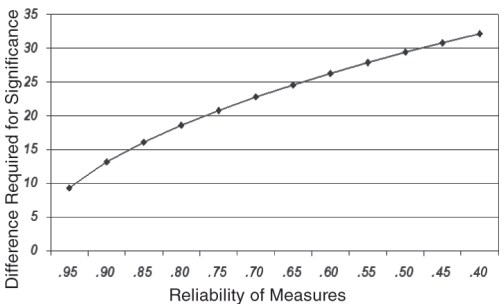

FIGURE 2.2. Relationships between reliability and the differences needed for significance when one is comparing two scores. Note that this figure assumes two variables with the same reliability and an *SD* of 15 at the 95% level of confidence.

in research should be questioned. This is particularly important in correlational research, because the extent to which two variables correlate is influenced by the reliability of each variable. Clinicians are advised not to use measures that do not meet reliability standards, because there will be too much error in the obtained scores to allow for reliable interpretation. This is especially important because the decisions clinicians make can have significant and longlasting impact on the lives of examinees.

Validity

Although reliability is important, reliable measurement of a construct with little validity would be of limited utility to the clinician and researcher. Validity is described as the degree to which empirical evidence supports an interpretation of scores that represent a construct of interest. For example, a measure of ASD should contain carefully crafted questions that accurately reflect the disorder. Researchers who study ASD and authors who develop tools to be used during the diagnostic process are especially burdened with the responsibility to carefully and clearly define the behaviors associated with these disorders. When the behaviors and characteristics associated with a disorder are thoroughly operationalized, then further development of the dimensions or factors that can be used for diagnosis may be clarified. This depends, of course, on the extent to which the items have adequate reliability.

Given the fact that methods for evaluating ASD, as well as our understanding of the underlying aspects of these disorders, are evolving, we have a particular responsibility to provide validity evidence of the effectiveness of any method we choose (rating scales, tests, interviews, etc.). This is not as simple a task as demonstrating reliability, because validity is harder to demonstrate and the findings will be directly related to the content of the tools used to study ASD, as well as the methodology employed. For example, the items included in a rating scale define and limit the scope of the information that is obtained. This can provide a broad or truncated view of the behaviors associated with a disorder. Choosing the standard against which measures are validated is also not foolproof, because today's so-called diagnostic gold standard (i.e., the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* [DSM-5]; American Psychiatric Association, 2013) will undoubtedly evolve to reflect future research findings. Similarly, research methodology is also important, particularly when typical children are being compared with those who have ASD. Special attention should be made to ensure that research findings provide a sufficient number of control groups to determine how those with ASD differ from typical children, as well as from those with other types of disorders.

In summary, the very nature of our understanding of ASD is influenced by the psychometric quality of the tests and methods we use to study these disorders, as well as by the selection of variables we use in our research. Clinicians should be mindful, however, that until there is sufficient maturity in the scope and quality of the instruments used during the diagnostic process, a good understanding of the strengths and weaknesses of all the methods used is necessary. This includes a careful understanding of the manner in which any measure of ASD is constructed.

Development of Scales to Assess ASD

There is a need for a number of well-standardized measures of ASD that have demonstrated reliability and validity. At this writing, there are several behavior rating scales that have been used in both applied and research settings, as well as structured clinical interviews and direct assessments that have varying degrees of reliability and validity. This amplifies the need for practitioners and researchers to have a good understanding of the psychometric qualities and standardization samples associated with these methods. Researchers and practitioners should also be informed about the development of any scale used to aid in the diagnosis of ASD; the test's development should be carefully described by its authors. Development of any scale should follow a series of steps to ensure the highest quality and validity. The development of tools to help diagnose ASD is a task that demands well-known procedures amply described by Crocker and Algina (1986) and by Nunnally and Bernstein (1994). These are now summarized.

Initial test development should begin with a clear definition of the behaviors that represent autism and other ASD. These behaviors and other defining characteristics must be written with sufficient clarity that they can be assessed reliably over time and across raters. Behaviors should be included that represent the characteristics that define children with autism or other ASD as completely as possible, are specific to these disorders, and reflect current conceptualizations of the disorders (such as the behaviors included in DSM-5). Definitional clarity is *required* for good item writing.

The next step is to develop an initial pool of questions, followed by pilot testing of the items. Pilot tests are designed to evaluate the clarity of the instructions and items, as well as the structure of the form and other logistical issues. For instance, it is important to be cognizant of the ways items are presented on the page, size of the fonts, clarity of the directions, colors used on the form, position of the items on the paper, and so forth. Analyses of reliability and validity are typically not of interest at this point, because sample size usually precludes adequate examination of these issues. Instead, the goal of pilot testing is to answer essential

questions such as these: Does the form seem to work? Do the users understand what they need to do? Are the items clear? Can the rater respond to each question?

In contrast, conducting experiments with larger samples that allow for an examination of the psychometric qualities of the items and their correspondence to the constructs of interest is the next important step. This effort is repeated until there is sufficient confidence that the items and the scales have been adequately operationalized. In each phase of the process, experimental evidence within the context of the practical demands facing clinical application should guide development, but some essential analyses such as the following should be conducted:

- Means and *SD*s, and *p* values (if dichotomous items are used), should be obtained for each item.
- Items designed to measure the same construct should correlate with a total score obtained from the sum of all those items designed to measure that same construct. If the correlations are low, their inclusion in the scale should be questioned.
- The contribution each item makes to the reliability of the scale(s) on which it is placed should be evaluated.
- An item designed to measure a particular construct should correlate more strongly with other items designed to measure that same construct than with items designed to measure different constructs. If this is not found, the item may be eliminated.
- The internal reliability of those items organized to measure each construct should be computed, as should the reliability of a composite score.
- The factor structure of the set of items may be examined to test the extent to which items or scales form groups, or factors, whose validity can be examined.

The procedures used at this phase are repeated until the scale is ready for standardization. The number of times these activities are repeated depends upon the (1) quality of the original concepts, (2) quality of the initial pool of items, (3) quality of the sampling used to study the instrument, and (4) consistency of the results that are obtained. The overall aim is to produce an experimental version of an instrument that is ready to be subjected to a larger-scale and more costly national standardization study. This would include sufficient data collection efforts to establish the reliability and validity of the final measure. Standardization requires that a sample of persons who represent the population of the country in which the scale will be used are administered the questions in a uniform manner, so that

normative values can be computed. Standardization samples are ordinarily designed to be representative of the normal population, so that those that differ from normality can be identified and the extent to which they differ from the norm (50th percentile) can be calibrated as a standard score to reflect dispersion around the mean. Development of norms is an art as much as a science, and there are several ways in which this task can be accomplished (see Crocker & Algina, 1986; Nunnally & Bernstein, 1994; Thorndike, 1982). The next tasks at this stage are collection and analysis of data for establishing reliability (internal, test–retest, interrater, intrarater) and validity (e.g., construct, predictive, and content). Of these two, validity is more difficult to establish and should be examined by using a number of different methodologies, with emphasis on assessing the extent to which the scale is valid for its intended purposes.

There are many different types of validity, making it impossible for validity to be determined by a single study. According to the *Standards for Educational and Psychological Testing* volume (American Educational Research Association, American Psychological Association, & National Council on Measurement Education, 2014), evidence for validity is required to support interpretations of test scores for intended use. Additionally, validity is noted to be an open-ended process with evidence collected prior to initial use as well as further data analysis as the test is in operational use. The standards relating to validity issues are to be addressed by authors and test development companies. Some of the more salient issues include the need to provide evidence that supports the following:

- Interpretations based on the scores the instrument yields.
- The appropriate relationships between the instrument's scores and one or more relevant criterion variables.
- The utility of the measure across a wide variety of demographic groups, or its limitations based on race, ethnicity, language, culture, and so forth.
- The expectation that the scores provided differentiate between groups as intended.
- The alignment of the factorial structure of the items or subtests with the scale configuration provided by the authors.

There is wide variation in the extent to which test authors document the development, standardization, reliability, and validity of their measures in test manuals. Some manuals provide sufficient descriptions that bring out the strengths of the scale; others provide limited details. Readers interested in illustrative manuals might look at those developed for the Universal Nonverbal Intelligence Test (Bracken & McCallum, 1997); the

Kaufman Assessment Battery for Children, Second Edition (Kaufman & Kaufman, 2004); and the Cognitive Assessment System, Second Edition (Naglieri et al., 2014). These examples illustrate how to provide detailed discussion of the various phases of development, as well as instructions about how the scores should be interpreted for the various purposes for which the measures were intended.

Documentation of development may end with the writing of the sections in the manual that describe the construction, standardization, and reliability/validity of the instrument, but authors also have the responsibility to inform users about how the scores should be interpreted (American Educational Research Association, American Psychological Association, & National Council on Measurement Education, 2014). This includes how test scores should be compared with one another, and authors should especially provide the values needed for significance when the various scores a measure provides are compared. This information is critically important if clinicians are to interpret the scores from any instrument in a manner that is psychometrically defensible.

Researchers and clinicians have a responsibility to choose measures that have been developed according to the highest standards available, because important decisions will be made on the basis of the information these measures provide. We suggest that for a scale to be considered acceptable for clinical practice, in addition to being reliable, it must have a standardized administration and scoring format with norms based on a large sample that represents the country in which the scale is used. This includes ample documentation of methods used to develop the measure, as well as ample evidence of validity and explicit instructions for interpretation of the scores that are obtained.

Obtaining information about the psychometric characteristics of instruments that could be used as part of the diagnostic process is a time-consuming and sometimes confusing task. Manuals provide different types of information; sometimes the information is clear and concise, and at other times it is hard to ascertain enough details to fully evaluate the results being presented. Comparisons across instruments are complicated by this inconsistency and by the logistical task of collecting the information. In the next section, we provide a systematic examination of the scales used to assess the behaviors associated with ASD. Our goal is to be informative about the specific details associated with important issues, such as reliability, validity, and standardization samples. The discussion of each test includes a general description of the scale, as well as reliability and validity information provided by the authors of these instruments in their respective test manuals. We end this chapter with a commentary on the relative advantages of these scales.

DESCRIPTIONS OF SCALES USED TO ASSESS ASD

Autism Diagnostic Observation Schedule, Second Edition

Description

The Autism Diagnostic Observation Schedule, Second Edition (ADOS-2; Lord, Rutter, et al., 2012; Lord, Luyster, Gotham, & Guthrie, 2012) is a semistructured assessment of communication, social interaction, restricted and repetitive behaviors, and play/imaginative interaction in children or adults suspected of having ASD. A referred individual is assessed with one of the five modules contained in the ADOS-2. Each module can be administered in 40–60 minutes and is geared for a child or adult at a particular developmental and expressive language level. Each module consists of a variety of standard activities and materials that allow the examiner to observe an individual engaging in behaviors typical of persons with ASD within a standardized setting in order to aid diagnosis.

The Toddler Module and Module 1 are most appropriate for children who are preverbal or do not consistency use phrase speech (i.e., flexible use of non-echoed, three-word utterances, and spontaneous, meaningful word combinations). The Toddler Module is appropriate for children ages 12–30 months, whereas Module 1 is appropriate for those children ages 31 months and older. The Toddler Module consists of 11 activities, whereas Module 1 consists of 10 activities that focus on the playful use of toys. Module 2 also focuses on the playful use of toys and contains 14 separate activities geared toward individuals at the phrase speech language level. Module 3 is intended for children and young adolescents who are verbally fluent and focuses on social, communicative, and language behaviors through 14 different activities. Finally, Module 4 consists of 10 mandatory activities and five optional activities that also examine social, communication, and language behavior through unstructured conversation, structured situations, and interview questions. This module is used in the assessment of verbally fluent older adolescents and adults.

Examiners take notes during ADOS-2 administration, and ratings are made immediately following the administration. Guidelines for ratings are provided in each module, and algorithms are used to formulate diagnosis. Separate algorithms are used for the interpretation of each module. For the Toddler and Modules 1–3, the ADOS-2 uses an overall cutoff score, which is the sum of selected items from the algorithm domains (Social Affect and Restricted and Repetitive Behaviors). The Social Affect domain includes items related to communication and reciprocal social interaction. The Restricted and Repetitive Behaviors domain includes items pertaining to observed restricted and repetitive behaviors. Module 4 uses separate cutoff scores for the Communication and Social Interaction total scores and the Communication + Social Interaction total score. Although items related

to stereotyped behaviors and restricted interests as well as imagination/creativity are presented on the form, they are not included in the algorithm for Module 4.

Although the ADOS-2 has many similarities to DSM-5 and the *International Classification of Diseases, Tenth Revision* (ICD-10; World Health Organization, 1992) models of diagnosing ASD, the ADOS-2 record form has a number of behaviors coded that are not included in the algorithms. In addition, the ADOS-2 does not include information about the age of onset or early history required for a DSM-5 or ICD-10 diagnosis. Last, the authors note, "In cases where ADOS-2 classification differs from the overall clinical diagnoses, clinical judgment should overrule the ADOS-2 classification in achieving a best-estimate clinical diagnosis" (Lord, Rutter, et al., 2012, p. 187). When overruling, clinicians need to specify how achieved scores do not adequately represent observed behaviors.

Description of the Comparison Group

The original validation sample of the ADOS consisted of 381 individuals, consecutively referred to the Developmental Disorders Clinic at the University of Chicago. The authors note the ADOS-2 items and codes are functionally identical to those of the original ADOS and that Module 4 was unchanged. Therefore, certain reliability and validity results from the ADOS continue to apply. As such, many of these individuals were also included in the ADOS-2 Extended Validation sample. The Extended Validation sample included 1,139 participants, with 325 participants having repeated ADOS assessments (between two and seven times), resulting in 1,620 assessments included in the sample. The Extended Validation sample was largely collected at the Developmental Disorders Clinic at the University of Chicago (n = 926). The remaining (n = 213) were obtained through a longitudinal study by the Treatment and Education of Autistic and Related Communication Handicapped Children (TEACCH) centers at the University of North Carolina, Chapel Hill, and University of Chicago, or in the University of Michigan Autism and Communication Disorders Clinic research studies. Last, the ADOS-2 replication sample (N = 1,259) was competed as a means to replicate the ADOS-2 algorithms. This sample was collected from 11 sites throughout the United States and Canada; data collected from the University of Michigan were excluded to develop an independent sample. In this sample, 23 participants had repeated assessments, all of which were included in the analysis.

For the three studies, samples were described in terms of (1) ASD, which encompasses all ASD diagnoses (e.g., autistic disorder, Asperger's disorder, pervasive developmental disorder not otherwise specified [PDD-NOS]); (2) autism (i.e., the more narrow diagnoses of autistic disorder; (3) non-autism

ASD (i.e., ASD diagnoses other than autism); and (4) non-spectrum (any other diagnosis that is not on the autism spectrum, such as language disorder or oppositional defiant disorder). It should be remembered the definitions used were based on *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision* (DSM-IV-TR; American Psychiatric Association, 2000).

The composition of the ADOS-2 Validation and Replication samples for each module are as follows:

- *Module 1—Few to No Words, Nonverbal Mental Age ≥ 15 Months*
  - Validation: autism (n = 69), non-autism ASD (n = 20), and non-spectrum (n = 16)
  - Replication: autism (n = 50) and non-spectrum (n = 5)
- *Module 1—Few to No Words, Nonverbal Mental Age ≥ 15 Months*
  - Validation: autism (n = 306), non-autism ASD (n = 51), and nonspectrum (n = 33)
  - Replication: autism (n = 245), non-autism ASD (n = 6), and non-spectrum (n = 46)
- *Module 1—Some Words*
  - Validation: autism (n = 201), non-autism ASD (n = 75), and non-spectrum (n = 76)
  - Replication: autism (n = 183), non-autism ASD (n = 21), and nonspectrum (n = 64)
- *Module 2—Younger Than 5 Years*
  - Validation: autism (n = 58), non-autism ASD (n = 49), and non-spectrum (n = 30)
  - Replication: autism (n = 53), non-autism ASD (n = 17), and non-spectrum (n = 18)
- *Module 2—Age 5 Years or Older*
  - Validation: autism (n = 126), non-autism ASD (n = 36), and nonspectrum (n = 30)
  - Replication: autism (n = 100), non-autism ASD (n = 9), and non-spectrum (n = 8)
- *Module 3*
  - Validation: autism (n = 129), non-autism ASD (n = 186), and nonspectrum (n = 83)
  - Replication: autism (n = 339), non-autism ASD (n = 45), and non-spectrum (n = 73)
- *Module 4*
  - Validation: autism (n = 16), non-autism ASD (n = 14), and non-spectrum (n = 15)
  - Replication: not completed, as no changes were made from original ADOS

For each sample and module, information within each diagnostic group of each module was further broken down by gender, chronological age, verbal mental age, and nonverbal mental age (this information is available in the test manual). The authors further report that the ethnicities of the participants for both studies were variable across modules and are provided in a range across groups. The ADOS-2 Validation sample ethnicity groups were as follows: white (71–91%), African American (4–27%), Asian American (1–5%), American Indian (0–0.8%), biracial (0–2.2%), and other or unknown race (0–0.6%); 3.4% of white participants were identified as Hispanic. The ADOS-2 Replication sample ethnicity groups were as follows: white (81–91%), multiracial (5–9%), African American (1.5–4%), Asian American (1–4%), American Indian (1%), and other or unknown race (1%); 2–7% of participants were identified as Hispanic.

The sample for the Toddler Module was collected on a total of 182 children ages 12–30 months who were seen for 360 assessments. The ASD group was seen between one and 14 times (mean [*M*] = 3.24, *SD* = 3.48) and the non-spectrum and typically developing groups were seen between one and 12 times (*M* = 2.43, *SD* = 2.68; *M* = 1.29, *SD* = 1.26, respectively). Children who were seen multiple times were participating in larger longitudinal studies. This sample consisted of children with ASD (*n* = 46), non-spectrum disorders (i.e., expressive language disorder, mixed receptive–expressive language disorder, nonspecific intellectual disability, and Down syndrome; *n* = 37), and typical development (*n* = 99). In this group, 76% were male and 24% female. Ethnicity groups for the Toddler sample were as follows: white (80%); Hispanic (4%); African American (2%); Asian American (2%); American Indian (1%); biracial, multiracial, or other (10%); and 1% did not report ethnicity. Information is further broken down by chronological age and socioeconomic status as measured by highest level of maternal educational attainment (this information is available in the test manual).

Reliability

In order to evaluate the reliability of individual items on the ADOS, the test authors obtained interrater reliability information for each module. As the authors purport the ADOS-2 items are functionally identical to the original ADOS items, reliability information from the original ADOS manual still apply. For Module 1, interrater reliability had a mean exact agreement of 91.5%, and all items had more than 80% exact agreement across raters. With the exception of items describing repetitive behaviors and sensory abnormalities, the mean weighted kappa coefficients exceeded .60 (*M* = .78). Items describing repetitive behaviors and sensory abnormalities were less frequently scored as abnormal within the autistic sample and proved

more difficult to score. One item, "behavior when interrupted," was eliminated due to poor reliability.

The mean agreement for Module 2 items was 89%, and all items exceeded 80% agreement. Out of the 26 items, the kappa value for 15 items exceeded .60 (*M* = .70), and the kappa value for the remaining items equaled .50, with the exception of four items: "unusual sensory interest in play material/person," "unusually repetitive interests or stereotyped behaviors," "facial expressions directed to others," and "shared enjoyment in interaction" had kappa values ranging from .38 to .49, with agreements from 78 to 93%. These items were either edited or eliminated due to poor reliability.

For Module 3, the mean exact agreement was 88.2%. Many items (17) had kappa values of .60 or better (*M* = .65), and all but two items received 80% or more agreement. The item "stereotyped/idiosyncratic use of words or phrases" was rewritten, and "communication of own affect," "social distance," "pedantic speech," and "emotional gestures" were either eliminated or collapsed within another item due to poor reliability.

Module 4 had a minimum of 80% exact agreement, with kappa coefficients exceeding .60 for 22 of the items (*M* = .66), and the remaining items having kappa values of .50 or higher. "Excessive interest in or references to unusual or highly specific topics or objects or repetitive behavior" had a kappa value of .41, and "responsibility" had a kappa value of .48. The items were kept because the agreement for both equaled 85%. "Attention to irrelevant details" and "social disinhibition" were eliminated due to poor reliability.

Interclass correlations for Modules 1–3 were computed from a subsample of the ADOS-2 Validation sample that allowed for the algorithm subtotals and totals for each module and the combined modules. When data were collapsed across all modules, the interclass correlation coefficients for the Social Affect domain was .96, Restricted and Repetitive Behaviors was .84, and overall total had a coefficient of .96. The authors note that these calculations were not made for Module 4, as no changes were made. Interrater agreement for diagnostic classification for autism versus nonspectrum was examined. For Module 1, agreement was 95%; for Module 2, agreement was 98%; and for Module 3, agreement was 92%.

Test–retest reliability for the ADOS-2 was obtained from 87 participants who were administered the same ADOS module two times within an average of 10 months. The authors describe this sample as primarily young, high-functioning children. For Modules 1–3, reliability for Social Affect ranged from .81 to .92, for Restricted and Repetitive Behaviors from .68 to .82, and overall total from .83 to .87.

Finally, the Toddler Module reliability of individual items demonstrated out of 38 items, 30 items had weighted kappa values equal to or greater than

.60, whereas the remaining had weighted kappa values greater than .45. Interrater item reliability was accessed across five categories that included Language and Communication, Reciprocal Social Interaction, Play, Stereotyped Behaviors and Restricted Interests, and Other Behaviors. Of 41 items, 39 demonstrated exact agreement of 80% or greater, whereas the rest held at 71% or greater. The reliability of domain scores and algorithm classification correlations were generated for both the All Younger/Older with Few to No Words (n = 10) and Older with Some Words (n = 4) algorithms. For the combined algorithm groups, interclass correlations were .95 for overall total, .95 for Social Affect, and .90 for Restricted and Repetitive Behaviors. Correlations are provided for both algorithm types in the manual.

The test–retest reliability for the Toddler Module was assessed using 39 children who received two administrations within a 2-month span. The sample included 17 children with ASD, 11 children with non-spectrum disorders, and 11 children who were typically developing. The test–retest reliability for the All Younger/Older with Few to No Words (n = 31) algorithm demonstrated correlations of .83 for Social Affect, .75 for Restricted and Repetitive Behaviors, and .86 for overall total. The absolute mean difference across the two assessments was 1.29 points (SD = 3.55) for the total score and .90 points (SD = 3.14) and .39 points (SD = 1.54) for Social Affect and Restricted and Repetitive Behaviors, respectively. The Older with Some Words (n = 8) algorithm demonstrated interclass correlations for test–retest reliability of .94 for Social Affect, .60 for Restricted and Repetitive Behaviors, and .95 for overall total. The absolute means difference for this algorithm across the two assessments was .63 points (SD = 2.13) for the overall total score, and .38 points (SD = 2.77) and .25 points (SD = 1.04) for Social Affect and Restricted and Repetitive Behaviors, respectively.

Validity

The authors of the ADOS-2 have provided results from factor-analytic studies of their scale. They reported that items from the Social Interaction and Communication domains loaded highly on the first factor, and a second factor consisted of items dealing with speech and gesturing. Factor-analysis studies of ADOS-2 Expanded Validity were similar to those of the original ADOS. Social Affect and Restricted and Repetitive Behaviors factors were positively correlated. Comparisons of children with autism, those with PDD-NOS, and those not on the spectrum are provided in the manual for each ADOS module. Typically, children with autism earned significantly higher scores on those items included in the modules than those with PDD-NOS, and the lowest scores were obtained by those not on the spectrum. The sample sizes by module and by group were wide-ranging from a low of 14 to a high of 375.

A receiver operating characteristic (ROC) analysis with using the ADOS-2 Expanded Validity study for Modules 1–3 found sensitivity between autism and non-spectrum to range from 91 to 98% and specificity to range from 50 (Module 1—Few to No Words, Nonverbal Mental Age ≤ 15 Months) to 94%. When non-autism ASD was compared with non-spectrum, sensitivity ranged from 72 to 95%, while specificity ranged from 19 (Module 1—Few to No Words, Nonverbal Mental Age ≥ 15 Months) to 83%. When ROC analyses were also completed using the ADOS-2 Replication sample for Module 1—Few to No Words, Nonverbal Mental Age ≥ 15 Months; Module 1—Some Words; Module 2—Younger Than 5 Years; and Module 3, results differed. Sensitivity for autism and non-spectrum ranged from 82 to 94%, whereas specificity ranged from 80 to 100%. For the non-autism ASD group and non-spectrum group, sensitivity ranged from 60 to 95% and specificity ranged from 75 to 100%.

Comparisons of mean scores for the Toddler Module was highest for those with ASD, followed by non-spectrum and typically developing groups for both algorithm types. Sensitivity and specificity studies were conducted for each algorithm type and divided into "unique developmental groupings" and "all visits." Group sizes were wide-ranging from 11 to 153 participants. Overall sensitivity ranged from 83 to 91%, while specificity ranged from 86 to 94%.

Autism Diagnostic Interview, Revised

Description

The Autism Diagnostic Interview—Revised (ADI-R; Rutter, Le Couteur, & Lord, 2003) is an extended interview that produces information needed to diagnose autism and assist in assessing other ASD. The ADI-R consists of 93 questions focusing primarily on three domains: Language/Communication; Reciprocal Social Interactions; and Restricted, Repetitive, and Stereotyped Behaviors and Interests. This interview should be administered by an experienced clinician to an informant familiar with the assessed individual's behavior and development. The assessed individual must have a mental age of at least 2 years. The interview takes approximately 1½–2½ hours to complete.

The interviewer records and codes detailed responses to the 93 questions, using the interview protocol. The interviewer then scores the interview, using one or more of the five algorithm forms. Algorithms are used to code up to 42 of the interview items in order to produce formal and interpretable results. The algorithms consist of both Diagnostic and Current Behavior algorithms. Diagnostic algorithms are used for diagnosis and focus on the individual's developmental history at ages 4–5 years, whereas

Current Behavior algorithms reflect symptoms at the time of testing and can be used for treatment and/or educational planning.

Summary scores are calculated for each of four domains (Qualitative Abnormalities in Reciprocal Social Interactions; Qualitative Abnormalities in Communication; Restrictive, Repetitive, and Stereotyped Patterns of Behavior; and whether the manifestations of behavior were evident [i.e., before 36 months of age]) for the Diagnostic algorithms. Cutoff scores are then used to determine the presence of ASD. There is only one cutoff for ASD, rather than separate cutoffs for autism and ASD, as on the ADOS.

Description of the Comparison Group

The ADI-R comparison group was developed by administering the ADI-R to several hundred caregivers of individuals both with and without autism; the individuals' ages ranged from preschool to early adulthood. Interviews were conducted as initial clinical assessments and research evaluations. No further information is provided on this sample of several hundred.

Reliability

The ADI-R manual presents interrater and test–retest reliability coefficients. Weighted kappa values are provided for the behavioral items of the four Diagnostic algorithm domains. These coefficients are broken down by age and come from one of two studies. In a sample of 19 children 36–59 months of age, the weighted kappa coefficients ranged from .63 to .89. In a sample of 22 individuals ages 5–29 years, weighted kappa coefficients ranged from .37 to .95. Test–retest reliability coefficients are also presented from a study of 94 preschool children with a test–retest period of 2–5 months. Coefficients were provided for the behavioral items, including Reciprocal Social Interaction; Abnormalities in Communication; and Restricted, Repetitive, and Stereotyped Behaviors and Interests of the Diagnostic algorithm domains (excluding Age of First Manifestation). Intraclass correlation coefficients ranged from .93 to .97.

Validity

The associations between the ADI-R and the Social Communication Questionnaire (SCQ; Rutter, Bailey, & Lord, 2003; see the description of that instrument later in this chapter), which is essentially a short form of the ADI-R, were examined for a sample of children with developmental language disorders to assess concurrent validity (Bishop & Norbury, 2002). The ADI-R was scored to distinguish those students meeting full DSM-IV/ICD-10 criteria for autism (this applied to eight out of a total sample of 21

children and eight out of the 14 with ASD), as well as those qualifying for a broad designation of ASD (children meeting criteria for two out of the three domains). Of the eight children meeting the full criteria on the ADI-R, six children scored 15 or more on the SCQ. Intercorrelations between the ADI-R and SCQ for the three ADI-R domains were examined. The Reciprocal Social Interaction domain had a Pearson correlation of .92; the Language/Communication domain correlation was .73; and the Restricted, Repetitive, and Stereotyped Behaviors and Interests domain correlation was .89. Within the ADI-R and SCQ, the cross-correlations between the Reciprocal Social Interaction and Language/Communication domains were .77 for the SCQ and .70 for the ADI-R. The Restricted, Repetitive, and Stereotyped Behaviors and Interests correlations with the other two domains were .48 and .53 for the SCQ, and .41 and .54 for the ADI-R.

Item-by-item agreement between the ADI-R and SCQ was provided. The ADI-R items were classified as present if a score of 1, 2, or 3 was obtained, whereas a score of 1 indicated agreement on the SCQ. Agreement between the items on the two tests ranged from 45 to 85%, with an average of 70.8%.

Autism Spectrum Rating Scale

Description

The Autism Spectrum Rating Scale (ASRS; Goldstein & Naglieri, 2009) is an observer-completed rating scale designed to aid in the diagnosis of individuals who may have ASD. The ASRS is completed by parents (or similar caregivers) or teachers (or similar professionals) who rate behaviors characteristic of children ages 2–5 years (Early Childhood version) and older children ages 7–18 years (School-Age form). All forms ask the rater to consider behaviors during the past month. The items measure behaviors characteristic of ASD and are organized to yield both empirically and rationally defined scales. There are three empirically derived scales (SelfRegulation, Social/Communication, and Unusual Behaviors) and an ASRS total scale. In addition to the factorially derived scales, there are several scales developed on the basis of locally organized item groups: Adult Socialization, Attention, Behavioral Rigidity, Emotionality, Peer Socialization, Language, Sensory Sensitivity, and Unusual Interests. The score for each of these scales is a *T*-score with a normative mean of 50 and an *SD* of 10. In addition, a short Screening version of the ASRS is provided, consisting of 15 items.

The authors state that the ASRS was developed to measure ASD and autismrelated problems, in order to allow clinicians to compare an individual to norm-based expectations in an objective and reliable manner. Because the ASRS items are linked to DSM-IV-TR symptoms of autistic

disorder, Asperger's disorder, and PDD-NOS, the information provided can also facilitate the process of differential diagnosis. Used in combination with other assessment information, results from the ASRS provide valuable information to guide diagnostic decisions. The results can also be used to help form individualized intervention plans and suggest behaviors to target in treatment, as well as to evaluate an individual's response to treatment. Finally, the 15-item ASRS Screening scale is intended to be used in large-scale prevention programs.

Description of the Comparison Group

The ASRS was standardized on a large sample of children and adolescents who were selected to be representative of the United States, with a proportional sample from Canada. Two samples of data were collected—one for the Early Childhood version and one for the School-Age version—to create norms for parent and teacher raters. Equal numbers of males and females, who ranged in age from 2 years, 0 months, through 18 years, 11 months, were included.

The normative sample included a total of 1,280 parent and 1,280 teacher ratings; 2,560 total ratings. The two- to five-year-old sample consisted of 640 ratings, completed by parents (n = 320) and teachers (n = 320). The 6- to 18-year-old sample consisted of 1,920 ratings completed by parents (n = 960) and teachers (n = 960). Sample characteristics were based on the 2000 U.S. Census report. Race/ethnicity backgrounds for the ASRS Parent Normative samples included white (61%), Hispanic (15.6%), African American (14.8%), multinational/other (4.4%), and Asian (4.1%), whereas the ASRS Teacher Normative samples consisted of white (59.2%), Hispanic (16.1%), African American (14.9%), multinational/other (5.4%), and Asian (4.4%). The manual also provides further race/ethnicity backgrounds for each of the forms and by gender. Data for parent education level was collected only on parent forms and consisted of less then high school (10.8%), high school graduate (26.6%), some college (29.1%), and college or higher (33.3%), with .03% missing. Geographically, the ASRS Parent Normative data were collected from the South (31.3%), Northeast (24.5%), Midwest (21.4%), West (19.5%), and Canada (2.8%), with .4% missing. Similarly, the ASRS Teacher Normative data were collected from the Northeast (37.9%), South (23.5%), Midwest (22.2%), West (11.7%), and Canada (4.7%). The diagnostic disruption of children included in the 2- to 5-year-old sample consisted of those diagnosed with autism (0.6%), speech/language impairments (3.1%), and developmental delays (0.6%), totaling 4.4%. The 6- to 18-year-old diagnostic disruptions had those diagnosed with autism (0.5%), emotional disturbance (1.1%), attention-deficit/hyperactivity disorder (ADHD; 4%), and speech/language impairments (3%), providing a total of 8.7% of the sample having a clinical diagnosis.

Reliability

The internal reliability coefficients for the empirically based scales for the Early Childhood ASRS are as follows: Social/Communication (39 items; reliability of .94 and .95 for parents and teachers, respectively), and Unusual Behaviors (23 items; reliability of .91 for parents and .85 for teachers) with total scale reliability of .95 and .94 for parents and teachers, respectively. The internal reliability coefficients for the empirically based scales for the School-Age ASRS are as follows for parents and teachers, respectively: Social/Communication (19 items; reliability of .91 and .92), Unusual Behaviors (24 items; reliability of .93 for both parents and teachers), and Self-Regulation (17 items; reliability of .92 and .93) with total scale reliability of .97 for parents and teachers. The 15-item ASRS Screening scale's reliability coefficients are .87 and .90 for parents and teachers, respectively, on the Early Childhood version, and .89 and .90 for parents and teachers, respectively, on the School-Age form.

Validity

Scores from the general population were compared with those diagnosed with ASD and other clinical diagnoses that included anxiety disorders, depressive disorders, and language disorders. A separate group of those diagnosed with ADHD was also compared. The group for the Early Childhood version parent form consisted of those from the general population (n = 133), ASD (n = 133), and other clinical diagnosis (n = 67), whereas the Teacher/Childcare Provided consisted of those from the general population (n = 123), ASD (n = 122), and other clinical diagnosis (n = 72). The samples for the School-Age parent form consisted of those from the general population (n = 211), ASD (n = 211), ADHD (n = 122), and other clinical diagnosis (n = 52). The teacher form for the School-Age version consisted of those from the general population (n = 227), ASD (n = 228), ADHD (n = 147), and other clinical diagnosis (n = 69). The manual provides a breakdown of demographic characteristics for each of these groups. Mean score differences demonstrated those from the ASD group had values above the recommended cut score of 60, with the ASRS total score above 70 (more than 2 *SD* above the normative sample).

The authors assess the validity of the ASRS to evaluate the diagnostic efficiency. In comparing the ASD group with the general population, they found the overall correct classification rate to range from 89.4 to 91.4% for the total score on all forms with the ASRS scales ranging from 88 to 93.5% on the Social/Communication scale and 85.2 to 94.8% on the Unusual Behaviors scale. DSM-IV-TR scale ranged from 89.75 to 94.1%. Sensitivity for the total score ranged from 89.8 to 91.1%, whereas specificity ranged from 88.6 to 92.2%. The authors also provide efficiency statistics for the

positive and negative predictive power, as well as the false-positive and false-negative rates and kappa score for the total score, ASRS scales, and DSM-IV-TR scale (see manual).

The manual provides several studies comparing the ASRS with other measures of ASD. These studies compared the *T*-score of the ASRS with the standard scores from the Gilliam Autism Rating Scale, Second Edition (GARS-2: Gilliam, 2006); and Gilliam Asperger's Disorder Scale (GADS; Gilliam, 2001); as well as the raw score from the Childhood Autism Rating Scale, Second Edition (CARS-2; Schopler, Van Bourgondien, Wellman, & Love, 2010). Correlations between the ASRS and GARS-2 ranged from .76 to 83, whereas the GADS ranged from .63 to .76. Correlations between the ASRS and CARS ranged from .06 to 50. The authors also compared the ASRS with the Cognitive Assessment System (CAS; Naglieri, Das, & Goldstein, 2013), which is based on A. R. Luria's conceptualization of major brain functions. The results found those with high ASRS total scores had mean standard scores ranging from 92.1 to 98.8 on the Planning, Simultaneous, and Successive scales, whereas the Attention scale had a mean score of 83.

Construct validity was assessed using several analyses to determine the extent to which the ASRS items form clusters that represent broad categories associated with the ASD diagnosis. Exploratory factory analysis suggested a two-factor solution was most suitable for the Early Childhood version, whereas a three-factor solution was most suitable for the School-Age version. For the Early Childhood version the exploratory factor analysis grouped behaviors related to socialization and communication into one factor and stereotypical and repetitive behaviors into a second factor. The School-Age version found socializing and communication problems loaded together on one factor, stereotypical and repetitive behaviors on a second factor, and self-regulatory behaviors on a third factor. Exploratory factor analysis also demonstrated consistency across genders, race/ethnicities, and clinical status.

Childhood Autism Rating Scale, Second Edition

Description

The Childhood Autism Rating Scale, Second Edition (CARS-2; Schopler et al., 2010) comprises three forms. The Childhood Autism Rating Scale, Second Edition, Standard Form (CARS2-ST) is for those ages 6 and younger, with an estimated IQ of 79 and lower. The CARS2-ST was formerly titled CARS in the original 1998 version. The Childhood Autism Rating Scale, Second Edition, High Functioning (CARS2-HF) version is designed for those ages 6 and older, with an estimated IQ of 80 or higher. These are 15-item behavior rating scales developed to help identify children

with autism, to evaluate varying degrees of the disorder, and help determine whether a more comprehensive evaluation for ASD is warranted. The CARS-2 was also developed to differentiate children with autism from those with other developmental disorders, particularly those with moderate to severe intellectual disabilities. CARS-2 ratings are based on a clinician's observations or on parent report. Behaviors are rated on a scale of 1 (within normal limits), 2 (mildly abnormal), 3 (moderately abnormal), and 4 (severely abnormal for that age), based on a one- or two-sentence description of the behavior being evaluated. Item scores for the CARS2-ST and CARS2-HF are summed and categorized. Raw scores for the CARS2-ST that range between 15 and 29.5 are considered minimal-to-no symptoms of ASD, 30–36.5 is considered the range of mild to moderate ASD, and 37–60 is considered the severe ASD range. The CARS2-HF ranges are 15–27.5 for minimal symptoms of ASD, 28–33.5 for mild to moderate symptoms of ASD, and 34 and higher is considered the severe symptom range for ASD. There is also a CARS-2 Questionnaire for Parents or Caregivers (CARS2-QPC) version. The CARS2-OPC form is designed to assist in gathering information regarding behaviors related to ASD to help inform scoring of the CARS2-ST and CARS2-HF. The 15 items included in the CARS-2 are based on the diagnostic criteria from Kanner (1943), the nine dimensions by Creek (1961), Rutter's (1978) definition, criteria proposed by the National Society for Autistic Children (1978), and DSM-IV-TR.

Description of the Comparison Group

The CARS-2 comprises three samples: the original 1998 CARS sample (N = 1,606), the CARS2-ST Verification sample (N = 1,034), and the CARS2-HF Development sample (N = 994). The original CARS sample were children who were referred to the North Carolina TEACCH program. The 1998 manual noted this comparison group comprised a referred sample of children suspected of having autism who had CARS scores below 30 (46%). The remaining 54% were identified as having autism. About half of the sample with autism had CARS scores that fell in the mild to moderate range, and half met the criteria for severe autism. This sample consisted of 23% females and 72% males (5% were missing demographic information), whose racial background was white (62%), African American (28%), other (3%), and missing (7%). The authors describe the sample as predominantly from low socioeconomic levels, based on the Hollingshead–Redlich twofactor index. Almost one-fourth (23%) of the sample fell in the lowest socioeconomic category (V) identified by the index. The rest of the sample was distributed as follows: IV (29%), III (20%), II (8%), and I (8%); missing (12%). The sample was further described on the basis of IQ as follows: 52% ≤ 69, 12% = 70–84, and 10% = 85 and above, with 26% missing.

Finally, the children varied in age as follows: <6 years = 53%, 6–10 years = 30%, and 11 and above = 10%; missing = 6%. No data on the minimum or maximum ages of the children included in the sample, or other characteristics (e.g., parental education) were provided. The degree to which this sample represents a population of children with autism in the state of North Carolina or the country was not provided. Importantly, these data were collected from the late 1960s through the late 1980s, according to the CARS 1998 manual.

The CARS2-ST Verification sample (N = 1,034) consisted of those already diagnosed with autism in various clinical settings and collected from four demographic regions: South (43%), Northeast (29%), West (17%), and Midwest (11%). The authors' state that to maintain consistency with the demographics of individuals diagnosed with autism, the sample was 78% males and 22% females. Race/ethnicity backgrounds were white (60%), black/African American (16%), Hispanic/Latino (13%), Asian/Pacific Islander (7%), and other (4%). Socioeconomic status, provided as head of household years of education completed, were less than high school diploma (13%), high school graduate (34%), some college (13%), college graduate (20%), postgraduate (13%), and missing (7%). IQ descriptions for the sample were ≤70 (81%) and 80–85 (19%) and were intently skewed to represent those with lower cognitive functioning with the acknowledgment that ratings for higher-functioning individuals may also occur. Last, ages ranged from 2 to 36 years old, with age bands in years of 2–5 (30%), 6–10 (43%), 11–15 (20%), and 16–36 (7%). Information regarding methods previously used to diagnose autism was not provided, nor was the severity of autism symptoms. Although the U.S. Census figures for 2000 (U.S. Census Bureau, 2000) are provided, there are large discrepancies between the Verification sample and actual U.S. population.

For the CARS2-HF Development sample (N = 994), individuals had a verity of clinical diagnoses that included high-functioning autism (n = 248), Asperger's disorder (n = 231), PPD-NOS (n = 95), ADHD (n = 179), learning disorder (n = 111), and "other internalizing and externalizing clinical disorders" (n = 69). To verify an absence of symptoms on the CARS2-HF, a small group of general education (n = 21) and non-autistic special education (n = 40) students were included. This sample was collected from various clinical settings around the United States: South (48%), Midwest (21%), Northeast (20%), and West (10%). Similar to the CARS2-ST verification sample, the CARS2-HF consisted of 78% males and 22% females. Race/ethnicity backgrounds were white (73%), black/African American (14%), Hispanic/Latino (6%), Asian/Pacific Islander (3%), Native American (1%), and other (3%). Socioeconomic status, provided as head of household years of education completed, is broken into less than high school diploma (8%), high school graduate (30%), some college (15%), college graduate (17%),

postgraduate (14%), and missing (16%). Ages ranged from 6 to 57 years old with age bands in years of 6–10 (35%), 11–15 (41%), and 16–57 (24%). The authors state that all individuals had an estimated IQ of 80 or higher—however, further information regarding IQ ranges is not provided.

Reliability

The CARS-2 manual presents internal, test–retest, and interrater reliability coefficients. In the CARS2-ST Verification sample (N = 1,034), internal reliability was estimated at an alpha coefficient of .93 with item-to-total correlations ranging from .43 to .81 (M = .69). The CARS2-HF Developmental sample (N = 994) obtained an estimated alpha coefficient of .96; item-to-total correlations ranged from .53 to .88 (M = .79).

Interrater reliability for the CARS (1998 version) was provided for the CARS2-ST. For the CARS, two trained independent raters evaluated 280 cases in the Development sample and obtained an estimated correlation of .71 with item correlations ranging from .55 to .93 (median = .71). Interrater reliability for the CARS2-HF was examined from two trained independent raters for 239 individuals in the Development sample. A correlation of .95 was obtained for total scores, whereas item correlations ranged from .53 to .93 (median = .73). Weighted kappa estimates were also calculated with a median level of agreement of .73, with items ranging from .51 to .90.

Information regarding test–retest reliability was provided only from the CARS (1998 version) Development sample, although the manual cites several external studies. The manual states 91 individuals assessed approximately 1 year apart resulted in an indication for the scale's stability over time—however, the correlation is not provided. Information is provided between the second and third evaluations, which the authors state is to avoid the effects of improvement in autistic behaviors frequently seen between the first and second assessment due to intensive treatment effort. Correlations between the second and third evaluation resulted in a correlation of .88 ($p < .01$) with means of 31.5 and 31.9 for the second and third evaluation, respectively; a kappa coefficient of .64 was also obtained. However, a time frame is not provided for the length between the first, second, and third evaluation. No test–rest reliability is provided for the CARS2-HF or updated for the CARS2-ST.

Validity

The authors assess internal structure of item ratings with the CARS2-ST using the Verification sample and the CARS2-HF using the Development sample. A similar pattern was observed between forms. The CARS2-ST correlations for item ratings ranged from .37 to .77; total raw score to item

correlations ranged from .55 to .84. The CARS2-HF between-item rating correlations for individuals without ASD ranged from .57 to .85, total raw score to item correlations ranged from .76 to .92. The Autism sample ranged from .35 to .69, whereas total raw score to item correlations ranged from .66 to .85. Factor-analytic results for the CARS2-ST yielded two component factors accounting for 59% of the variance with the first factor related to communication and sensory issues and the second factor related to emotional issues. In a factor analysis of the CARS-HF for those with and without ASD, three factors accounted for 59% of the variances in the ratings. The first factor related to social and emotional issues, the second related to cognitive functioning and verbal ability, and the third related to sensory issues.

The authors provide a ROC analysis for the original CARS to identify those with and without ASD using the total raw scores and found a sensitivity value of .88 and specificity value of .86, resulting in a false-negative rate of 12% and false-positive rate of 14%. The authors report ratings obtained for the CARS2-ST Verification sample was consistent with the original findings. Several independent studies are provided in the manual that found similar results when a ROC analysis was conducted using the original CARS.

The Development sample of the CARS2-HF was examined for differences in total raw scores between diagnoses. They found those with high-functioning autism obtained a mean of 35.3 (SD = 6.9), followed by PPD-NOS (M = 33.6, SD = 7.2), Asperger's disorder (M = 32, SD = 6.1), mixed clinical (M = 24.8, SD = 7.7), ADHD (M = 19.6, SD = 5.1), nonverbal learning disorder (M = 19, SD = 6.4), learning disorder (M = 18.7, SD = 5.1), general education students (M = 17.3, SD = 2.1), and special education students (M = 17, SD = 2.6). The manual also provides mean scores for each item in these groups. A ROC analysis to distinguish the non-autistic and autistic groups found sensitivity to be .81 and specificity to be .87 with a corresponding false-positive rate of 11% and false-negative rate of 23%.

The authors assessed criterion-related validity for the original CARS by comparing total scores to clinical ratings obtained during the same diagnostic session (r = .85, p < .001). Total scores were also correlated with independent clinical assessments made by a child psychologist and a child psychiatrist. This was based on information obtained from referral records, parent interviews, and nonstructured clinical interviews (r = .80, p < .001).

The relationship between the CARS2-ST and CARS2-HF with other measures evaluating autism was compared. Comparing the CARS2-ST and the ADOS (n = 37) provided a correlation of .79, whereas comparing CARS2-HF and the ADOS (n = 76) provided a correlation of .77. Correlations between clinician-generated CARS-2 total scores and mothers' Social Responsiveness Scale (SRS) total scores (n = 293) resulted in a moderate

relationship with a correlation of .38 for the CARS2-ST and .47 for the CARS2-HF. The authors purport the moderate relationship found is typically seen between clinician and parent ratings of children's behaviors. Several independent studies compared the original CARS with other measures of autism. Pilowsky, Yirmia, Shulman, and Dover (1998) reported 85.7% diagnostic agreement between the ADI-R and original CARS in a study of 83 children and adults with an overall kappa of .36. Another study found significant Pearson correlations on the original CARS scores and ADI-R that ranged from .60 on the ADI-R Communication subscale to .81 on the Social Impairment subscale and the CARS total score. Last, several studies compared the CARS with the Ritvo–Freeman Real Life Rating Scale (Ritvo et al., 2000) and Autism Behavior Checklist (Eaves, Campbell, & Chambers, 2000) with varying results.

Psychoeducational Profile, Third Edition

Description

The Psychoeducational Profile, Third Edition (PEP-3; Schopler, Lansing, Reichler, & Marcus, 2005) is an instrument designed to evaluate cognitive skills and behaviors typical of individuals characterized as having ASD and other developmental disabilities. This instrument is appropriate for children between the ages of 6 months and 7 years, for the purposes of planning educational programming and in the diagnosis of autism and other ASD. The test manual outlines four specific purposes of the PEP-3: to identify an individual's strengths and weaknesses, to aid in diagnosis, to establish developmental and adaptive level, and to serve as a research tool.

The PEP-3 has two major components: the Performance Part and the Caregiver Report. The Performance Part is administered through direct observation and testing, and consists of 10 subtests (six measuring developmental abilities and four measuring maladaptive behaviors) that form three composite scores: Communication, Motor, and Maladaptive Behavior. The Caregiver Report is completed by a parent or caregiver based on daily observations of the child. The Caregiver Report consists of two sections: (1) child's current developmental level and (2) degree of problems in different diagnostic categories. This information can be used to aid in clinical diagnosis. The Caregiver Report contains three subtests: Problem Behaviors, Personal Self-Care, and Adaptive Behavior.

Items on the PEP-3 are scored according to scoring criteria provided in the Examiner Scoring and Summary Booklet. Normative data are provided to facilitate a normative analysis, which allows the examiner to establish adaptive/developmental levels and make comparisons of the child to other children with autism. These scores can also be used in clinical analysis and provide information on a child's passing, emerging, or failing performance

on individual items, as well as appropriate, mild, or severe performance on individual Maladaptive Behavior items.

Normative scores allow examiners to compare a child's developmental age to that of a typically developing sample. The test authors state that a child identified as having ASD characteristically has an uneven developmental profile in relation to the developmental subtests. This developmental profile can then be used for determining the child's strengths and weaknesses. Percentile ranks were determined based upon a comparison sample with ASD and are available for subtests (and composite scores for the developmental subtests). The manual provides interpretive guidelines for these scores. Percentile scores above 89 are considered to be at the adequate developmental/adaptive level, 75–89 at the mild level, 25–74 at the moderate level, and below 25 at the severe level. Percentile ranks for the Maladaptive Behavior composite can also be used in interpretation. The manual states that a score lower than the 90th percentile in this composite usually places a child on the autism spectrum. Scores on the Problem Behaviors and Adaptive Behavior subtests, as well as the Caregiver Report, can be used to reinforce this interpretation.

Description of the Comparison Group

A sample of 407 children with autism and other ASD, as well as 149 typically developing children, was used for the PEP-3 normative sample. In the group with ASD, 95% of the children were classified as having autism, 4% as having Asperger syndrome, and 1% as exhibiting a developmental delay. Children in the sample ranged from ages 2 to 21 years (2 years, n = 38; 3 years, n = 60; 4 years, n = 63; 5 years, n = 51; 6 years, n = 48; 7 years, n = 23; 8 years, n = 27; 9 years, n = 21; 10 years, n = 19; 11 years, n = 16; 12 years, n = 15; 13–21 years, n = 26). The sample closely matched the U.S. population with regard to geographic area, gender, race, Hispanic ethnicity, family income, and educational attainment of parents.

Individuals in the typically developing sample consisted of 149 children between the ages of 2 and 6 (2 years, n = 27; 3 years, n = 33; 4 years, n = 36; 5 years, n = 27; 6 years, n = 26). This sample was 53% female and 47% male. The normative population closely matched the U.S. population on the domains of geographic area, race, Hispanic heritage, family income, educational attainment, and disability status.

Reliability

Internal consistency was assessed in a sample of individuals with autism at 11 age intervals (ages 2–11). Average alpha coefficients for Performance Part subtests, Caregiver Report subtests, and composites ranged from .84

to .99. Coefficient alphas were also provided for six subgroups of individuals with autism and the normally developing sample. The six subgroups, and the range of their alpha values for the Performance Part subtests, Caregiver Report subtests, and composites, were as follows: white (.78–.99), black (.76–.99), other race (.80–.99), Hispanic (.79–.99), male (.77–.99), female (.81–.99), and the normally developing sample (.75–.97).

Test–retest reliability was also examined in a sample of 33 children with autism between the ages of 4 and 14 residing in California, Oklahoma, and Texas. The sample consisted of 28 males and five females, and was also broken down by race and Hispanic ethnicity (white = 24, black = 4, other race = 5, Hispanic ethnicity = 6). The correlation coefficients ranged from .94 to .99 for both Performance Part subtests and Caregiver Report subtests. Correlation coefficients could not be calculated for composite scores, as raw data were used. The time lapse between the first and second test was 2 weeks.

Interrater reliability was assessed by using polychoric correlations, because items on the Caregiver Report are ordered categorical data. The sample used in this reliability study consisted of 40 individuals ages 2–10 from seven different states. Of the 41 participants, 33 were male and seven were female, 32 were white, six were black, two were of other races, and one was Hispanic. Nine of the 40 children did not have a disability, 29 were diagnosed with autism, and two were diagnosed with Asperger syndrome. Two parents of each child independently completed the Caregiver Report, and polychoric correlations for the items on the Problem Behaviors, Personal Self-Care, and Adaptive Behavior subtests were computed. Polychoric correlations for the items on the Adaptive Behavior subtest ranged from .70 to .91 (M = .85), Personal Self-Care item correlations ranged from .65 to 1.00 (M = .90), and correlations for the Adaptive Behavior subtest items ranged from .52 to .90 (M = .78). It should be noted that one item on the Adaptive Behavior subtest was eliminated because it had a very low correlation.

Validity

Median item discrimination coefficients were calculated by the test authors for a sample of children with autism ages 2–12 to assess the degree to which an item would correctly differentiate among test takers. Such coefficients were calculated for 11 age intervals for each subtest of the Performance Part and the Caregiver Report. Item difficulty coefficients were also calculated at these 11 age intervals to determine the items that were too easy or too difficult and arrange them in order from least to most difficult.

In order to detect differential item function (DIF), a logistic regression procedure was applied to all PEP-3 subtest items. The sample of individuals

with autism was used to make comparisons between these groups: male versus female, black versus non-black, and Hispanic versus nonHispanic. Four of these comparisons were found to illustrate DIF at the .001 significance level. However, after reviewing these items, the test authors suggested that the four items exhibited benign DIF.

Criterion prediction validity was assessed in four studies by examining the relationship between the PEP-3 and four criterion measures. First, the authors examined the relationship between the PEP-3 and the original Vineland Adaptive Behavior Scales, Expanded Form (VABS; Sparrow, Balla, & Cicchetti, 1984) for a sample of 45 children with autism between the ages of 2 and 14. In general, the correlations were high, with only a few exceptions (e.g., Vineland Motor Skills with PEP-3 Problem Behaviors). The second study (N = 68) examined the correlations between the CARS and the PEP-3. Significant and large correlations were found. Similarly, the third study involved the correlations of the PEP-3 with the Autism Behavior Checklist, Second Edition (Krug, Arick, & Almond, 2008). The results for this sample of 316 children suggested that the two scales are highly correlated.

The test authors calculated correlations between all subtests and found that these correlations ranged from .39 to .90, with a mean of .68. The authors state that coefficients for the subtests range from moderate to very large, and that the mean coefficient falls within the large range. Because of this, they suggest that the PEP-3 subtests measure different skills or behaviors and that evidence thus exists for construct identification validity. These intercorrelations were further subjected to confirmatory factor analysis to test the degree to which the subtests' assignment to the three composites were supported by data from the standardization sample. The results indicated that the three composites (Communication, Motor, and Maladaptive Behavior) could be considered a viable structure for this instrument.

Social Communication Questionnaire

Description

The SCQ (Rutter, Bailey, et al., 2003) is a 40-item rating scale completed by parents to assess the symptoms associated with ASD. The content of the scale is the same as that of the ADI-R (Rutter, Le Couteur, at al., 2003), reviewed above, with items worded identically, but it is administered as a parent questionnaire rather than via an extended interview. The scale uses a yes/no format and, according to the test manual, takes approximately 10 minutes to complete and 5 minutes to score. Raw scores are summed to yield a total score, which is interpreted based on the form being used and recommended cutoff scores. The SCQ has two forms: Lifetime and Current

Behavior. The Lifetime form assesses the individual's entire developmental history, whereas the Current Behavior form assesses behavior in the most current 3 months. The Lifetime form is considered more useful for diagnosing or screening ASD, while the Current Behavior form can be beneficial for developing treatment plans.

According to the authors, the SCQ has three main uses. First, it can be used as a screening device for the presence of ASD. If a child is suspected of having ASD after being screened, further clinical assessment should be conducted. The SCQ is an alternative to the ADI-R for use when time does not permit a lengthy assessment, such as in screening; the questions are identical, so one or the other can be used, but not both. The subscores produced by the SCQ can also be used to match the domains of the ADI-R (Reciprocal Social Interaction; Language/Communication; and Restricted, Repetitive, and Stereotyped Behaviors and Interests). Although the production of subscores can be used for interpretation, the manual warns that these subscores have not been adequately researched. A second use of the SCQ is for research purposes; it can be used with groups of children diagnosed with ASD to compare symptoms across groups. A third identified use of the SCQ is its ability to identify severity of ASD symptoms or changes in severity of symptoms over time. This is accomplished through the use of the Current Behavior form.

Description of the Comparison Group

Raw scores from the SCQ are compared with those earned by a sample of 200 children who had participated in previous studies using the ADI-R. The children in this sample had a variety of developmental disabilities: 83 had autism, 49 had atypical autism, 16 had Asperger syndrome, seven had fragile X syndrome, five had Rett syndrome, 10 had conduct disorder, seven had language delay, 15 had intellectual disability, and eight had other clinical diagnoses.

Reliability

Information is provided on the internal consistency of the SCQ as a measure of reliability. Alpha coefficients were computed in two different ways. First, a sample of 214 children with both ASD and non-spectrum diagnoses was divided into five different groups. These groups consisted of a "no-language" group and four "language" groups divided by age. Alpha coefficients for these groups ranged from .84 to .93. Next, internal consistency was examined by dividing the 157 children in the language group into one of three diagnostic categories: autism, other ASD, and non-spectrum. Measures of internal consistency for these groups ranged from .81 to .92.

Validity

Of the 39 items scored on the SCQ, 33 showed statistical differentiation of children with ASD from those with other abnormalities. The items that did not show differentiation primarily concerned abnormal language features. These items had a relatively high frequency among children without ASD, but correlated with the total score (.64, .53, .45, and .57). Two items (selfinjury and unusual attachment to objects) differentiated at the 7% significance level and showed more modest correlations with the total score (.37 and .27, respectively). Correlations were also calculated for the total score and domains (Reciprocal Social Interaction; Language/Communication; and Restricted, Repetitive, and Stereotyped Behaviors and Interests). All correlations were significant at the .0005 level within and across the domains and ranged from .31 to .71 (Berument, Rutter, Lord, Pickles, & Bailey, 1999).

Three- and four-factor solutions were explored for 39 items of the SCQ (items 2–40). Analysis suggested that a fourfactor structure appeared to be an acceptable fit. Principalcomponent factoring with varimax rotation yielded four factors and explained 42.4% of the total variation of the SCQ data; 24.3% (eigenvalue = 9.7) was accounted for by a social interaction factor (1), 8.7% (eigenvalue = 3.38) by a communication factor (2), 5% (eigenvalue = 1.94) by an abnormal language factor (3), and 4.5% (eigenvalue = 1.74) by a stereotyped behavior factor (4). The alpha reliability was .90 for the total scale, .91 for factor 1, .71 for factor 2, .79 for factor 3, and .67 for factor 4. The individual item-to-total scores were positive and mainly substantial, with a range of .26–.73. The four factors mapped onto the three domains that were operationalized by the ADI-R algorithm criteria. Factor 1 coincided with the Reciprocal Social Interaction domain; factor 4 coincided with the Restricted, Repetitive, and Stereotyped Behaviors and Interests domain; and the Language/Communication domain items were mainly divided between factors 2 and 3 (Berument et al., 1999).

ROC analysis and a series of *t*-tests were used to assess the discriminative power of the SCQ (Berument et al., 1999). After examining the area under the curve (AUC), the authors reported that the SCQ was able to differentiate ASD (including autism) from non-ASD conditions, including intellectual disability (AUC = .86). The SCQ also effectively differentiated between autism and non-ASD conditions other than intellectual disability (AUC = .94), autism and intellectual disability (AUC = .92), and autism and other ASD (AUC = .74), although this last distinction was less clear-cut.

Analyses were then repeated, using an SCQ score that did not include the six items that failed to differentiate the groups at the 5% significance level. The authors reported that some improvement in discriminative validity was obtained. However, the discriminative validity between autism and

other ASD was worse. The discriminative validity of the SCQ was then compared with that of the ADI-R. AUC results were contrasted for ASD versus non-ASD conditions (AUC = .88 and .87, respectively), autism versus intellectual disability (AUC = .93 and .96, respectively), and autism versus other ASD (AUC = .73 and .74, respectively).

The authors also reported that groups differed in IQ distribution, and considered that SCQ diagnostic differentiation could be due to this differentiation. In order to investigate this possibility, analyses were repeated within the identified IQ bands. Data came from various studies, and as a result, several different IQ tests were used to assess cognitive abilities. Results showed that in the comparison group, the SCQ score was the lowest (8.39) in the group with an IQ above 70 and highest in the group with severe intellectual disability (14.74), and that the SCQ score did not vary by IQ within the group with ASD. The diagnostic differentiation within the IQ bands was significant and clearest in the group with an IQ above 70.

Another set of analyses was conducted to examine whether individual behavioral domains of the SCQ provided better diagnostic information than that obtained with the total score. Items of the SCQ were placed in one of three domains determined by the equivalent items on the ADI-R. All three domains provided differentiation of ASD from other disorders (AUC ranged from .79 to .83), and differentiation on the total score was stronger (AUC = .90). The authors reported that the total score provided the best differentiation. This is supported by the finding that the Restricted, Repetitive, and Stereotyped Behaviors and Interests domain was not good at differentiating autism from intellectual disability (AUC = .70) or autism from other ASD (AUC = .59).

The authors reported that the ROC analysis for the total SCQ suggests a score of 15 or more as the cutoff for differentiating ASD from other diagnoses (sensitivity = .85, specificity = .75, positive predictive value = .55 for the sample). The authors also suggested that other cutoffs may be desirable for general population samples or other purposes. The cutoff of 15 or more had a sensitivity of .96 and a specificity of .80 for autism versus other diagnoses, excluding intellectual disability, and a sensitivity of .96 and a specificity of .67 for autism versus intellectual disability. Finally, a higher cutoff of 22 or more was required to differentiate autism from other ASD with a sensitivity of .75 and a specificity of .60.

As noted earlier, the relationships between the ADI-R and SCQ were examined in a sample of children with developmental language disorders to assess concurrent validity (Bishop & Norbury, 2002). See the "Validity" section in the ADI-R for the results of that study.

The relationships between the SCQ and ADI-R total scores were reported by the test authors for 81 children involved in an international genetics study. The correlation between the SCQ and ADI-R was .78 and

the intercorrelations among domains ranged from .44 to .77 ($p < .01$). The authors reported that the intercorrelations did not vary by age, gender, or IQ. When scores of 1, 2, and 3 on the ADI-R were compared with SCQ scores of 1, agreement ranged from 36.6 to 91.9%, with an average of 69.8%. When the ADI-R scores of 0 and 1 were collapsed and contrasted with scores of 2 on the SCQ, agreements were very similar and had an average of 68.5%.

Social Responsiveness Scale, Second Edition

Description

The Social Responsiveness Scale, Second Edition (SRS-2; Constantino & Gruber, 2012) comprises four forms that allow for evaluation from ages 2 years, 5 months, through adulthood and includes the Preschool form (ages 2 years, 5 months, to 4 years, 5 months), School-Age form (ages 4–18 years), and Adult Self-Report and Adult Relative/Other-Report forms (ages 19 years and older). The forms consist of 65-item questionnaires that cover various dimensions of interpersonal behavior, communication, and repetitive/stereotypical behavior. Items on the scale focus on the behavior of the child or adult, and the rater's responses are given in a Likert-type format. Each form takes approximately 15 minutes to complete and 5–10 minutes to score.

The test authors recommend that the SRS-2 results be used in different ways, depending upon the goal of assessment. When the SRS is used as a broad screening tool for any ASD in general populations, a cutoff raw score of 70 is recommended. When the instrument is used for screening in clinical and educational settings for children suspected of having social development problems, a cutoff score of 85 is recommended. For all forms, when SRS-2 scores are converted to *T*-scores, a value of ≤ 59 is within normal limits and not associated with clinically significant ASD; 60–65 is considered to be in the mild range, indicating deficiencies in reciprocal social behaviors and generally attributed to other clinical reasons (e.g., ADHD, language delays); scores from 66 to 75 are in the moderate range and represent deficiencies in reciprocal social behaviors that are often seen in those with moderate ASD and social (pragmatic) communication disorder; scores ≤ 76 are in the severe range and are strongly associated with a clinical diagnosis of ASD. Two scales, with proposed alignment at the time the manual was written to DSM-5 criteria, are provided and include Social Communication and Interaction and Restrictive Interests and Repetitive Behaviors. Last, the SRS-2 also yields *T*-scores for five treatment subscales: Social Awareness, Social Cognition, Social Communication, Social Motivation, and Restrictive Interests and Repetitive Behaviors. The authors note that

application of these treatment subscales should be limited to research of clinical investigation to target alleviation of symptoms as available evidence does not support individual interpretation of these subscales. Instructions for interpretation of DSM-5-compatible subscales and treatment subscales scores are provided in the test manual.

Description of the Comparison Group

The SRS-2 standardization included three independently collected samples. The School-Age standardization sample consisted of 1,014 children (ages 4–18) with 2,025 reports that included a parent/caretaker (n = 1,014) and teacher (n = 1,011) report for each child and was collected from four geographic regions: South (44.7%), West (23.2%), East (16.2%), and Midwest (15.9%). To maintain consistency with the 2009 U.S. Census (U.S. Census Bureau, 2009) figures, 51.2% were female and 48.7% were male. The racial composition of this normative group was as follows: white (59.5%), Hispanic/Latino (16.6%), black/African American (15.8%), Asian (5.7%), Native American (0.3%), and other (1.6%). Parental educational level included less then high school graduate (13.7%), high school graduate (26.1%), some college (24.8%), and 4 years of college or more (35.4%).

The Preschool standardization sample included 247 children with 474 reports that included a parent (n = 247) and teacher (n = 227) report for each child. Ratings were obtained when the children were between 30 and 54 months old with 58–70 children in each of the four age bands. The sample consisted of 51.4% male and 48.6% female from four geographic regions: South (29.6%), Midwest (27.5%), East (21.9%), and West (20.6%). Race/ethnic background consisted of white (64.4%), Hispanic/Latino (16.6%), black/African American (14.2%), Asian (1.6%), and other (2.2%). Educational level for parents included less then high school graduate (13.4%), high school graduate (32.4%), some college (23.5%), and 4 years of college or more (30%).

The Adult standardization sample involved 702 adults (ages 18–89) with 2,210 reports with at least three SRS-2 rating forms: Adult Self-Report and two Relative/Other-Report forms or three Relative/Other-Report forms if an Adult Self-Report form was not completed. Relative/Other-Report forms were completed by parents (n = 195), siblings or other relative (n = 519), spouse (n = 338), or close friend (n = 521). Of the sample, 54% were female and 46% were male. Geographic regions included South (37%), East (24.9%), Midwest (21.2%), and West (16.8%). The racial composition for the adult normative group was as follows: white (69.4%), black/African American (14.4%), Hispanic/Latino (13.8%), Asian (1.6%), Native American (0.1%), and other (0.7%). Educational attainment included less then high school graduate (14.5%), high school graduate (35.6%), some

college (16%), and 4 years of college or more (32.5%). Data for the Adult standardization sample were collected from public and private schools, as well as church and other groups that the authors purport to be a broad cross-section of the local community.

A clinical sample (N = 7,921) was obtained from the Interactive Autism Network Research Database at Kennedy Krieger Institute and John Hopkins–Baltimore, dated February 14, 2011. This sample included clinical subjects (n = 4,891) who held a formal ASD diagnosis prior to completing the SRS-2 and unaffected siblings (n = 3,030). Subjects ranged from 4 to 18 years old, with the sample being predominantly younger in age. In the clinical sample, 82.7% were males and 17.3% were females, whereas the unaffected siblings were 53.2% males and 46.8% females. Additionally, the combined sample were predominantly white (82.7%), followed by multiple (4.7%), Hispanic/Latino (4.5%), black/African American (2.7%), other (1.6%), Asian (1%), and Native American (0.1%). Geographic regions for the total sample included South (33.8%), East (24.1%), Midwest (24%), and West (18%). The manual does not provide statistics on educational attainment.

Reliability

Internal consistency, construct temporal stability, and interrater agreement data are provided in the SRS-2 manual. For the School-Age form, parent and teacher ratings were broken down by age and gender. The alpha coefficient for internal consistency on parent data for males was .95 (n = 493) and for females was .95 (n = 518), with age-based alpha coefficients ranging from .92 to .97. In turn, the alpha coefficient for internal consistency with teacher data for the male population was .96 (n = 509) and for females was .95 (n = 505) with aged-based coefficients ranging from .92 to .97. Interrater agreement between parent and teacher forms was .61 for males (n = 467) and .60 for females (n = 476), and aged-based coefficients ranged from .42 to .77.

Internal consistency alpha coefficients for the Preschool parent form were .93 for males (n = 127) and .95 for females (n = 120), and age-based coefficients ranged from .93 to .95. The teacher form alpha coefficients for internal consistency were .94 for males (n = 116) and .96 for females (n = 111), and age-based coefficients ranged from .93 to .95. Interrater reliability coefficients between parent and teacher forms was .77 for both males (n = 116) and females (n = 111), and age-based coefficients ranged from .70 to .78.

The Adult Self-Report form yielded internal consistency alpha coefficients of .95 for males (n = 288) and .94 for females (n = 349), whereas the Relative/Other-Report form yielded .97 for males (n = 732) and .96 for

females (n = 841). Age-based internal consistency alpha scores ranged from .93 to .96 for the Self-Report form and .94–.97 for the Relative/Other-Report form. Finally, interrater reliability coefficients were provided for self, mother, father, relative, spouse, and other. These ranged from .61 (self–other rater pairings) to .95 (other–mother pairings).

Internal consistency for the clinical sample yielded alpha coefficients of .95 for males and .94 for females for the clinical subjects, and .97 for both males and females of the unaffected siblings. Age-based internal consistency alpha scores for the clinical subjects ranged from .94 to .96 for males and from .96 to .97 for females of the unaffected siblings. For the clinical groups, internal consistency yielded alpha coefficients of .95 for all groups (i.e., autistic/autistic disorder, Asperger syndrome, PPD-NOS, PDD, and ASD).

Validity

The test authors used four independent samples to assess the factor structures of the SRS-2. These samples included the standardization subjects for the Preschool, School-Age, and Adult (Relative/Other-Report) forms, as well as a clinical sample. Results of the two-factor analysis of these samples indicated that a majority of items reflected autistic traits that fall under the Social Communication and Interaction scale, which comprises the treatment subscale items for Social Awareness, Social Cognition, Social Communication, and Social Motivation. The remaining items reflected a measure of the Restricted Interests and Repetitive Behaviors scale, which has a treatment subscale of the same name.

The authors report that a ROC analysis of the clinical sample revealed high degrees of sensitivity and specificity for the total raw scores. A total raw score of 60 was associated with a sensitivity of .93 and specificity of .91 for any ASD, whereas a total raw score of 75 was associated with a sensitivity of .84 and specificity of .94. The manual provides separate scores for males and females along with four total scores between 60 and 75.

In the manual, the authors cite more than 40 separate studies that represent over 30,000 independent administrations of the SRS-2 that evaluate assessment of autism and comorbid disorders, treatment effects, and validity with other instruments. The authors describe several independent studies from the United States and other countries that demonstrated similar findings in psychometric properties for mean differences, internal consistency, retest reliability and temporal stability, interrater reliability, and convergent validity, as well as sensitivity and specificity.

The authors describe a number of studies that investigated the comparison of the SRS-2 with other ASD-directed measures. The most widely investigated comparisons were between the SRS-2 with the SCQ. In studies

of mixed samples, correlations ranged from .50 to .65 (Bölte, Poustka, & Constantino, 2008; Bölte, Westerwald, Holtmann, Freitag, & Poustka, 2011; Charman et al., 2007; Pine, Guyer, Goldwin, Towbin, & Leibenluft, 2008; Granader et al., 2010). Charman et al. (2007) also found in children identified as developmentally at risk that the SCQ produced a sensitivity of .86 and specificity of .78, whereas the SRS-2 produced a sensitivity of .78 and specificity of .67. Analyses of these differences found the SCQ demonstrated a stronger performance in those with lower intellectual functioning (IQ below 75) and performed similarly in a higher IQ subgroup.

In comparing the SRS-2 with the Children's Communication Checklist (CCC), Pine et al. (2008) found correlations of $r = .49$ and .72 with the two ASD-focused subscales of the CCC, whereas Charman et al. (2007) found a correlation of $r = .75$. In comparing the two measures between children with ASD and those without ASD, Charman et al. (2007) found the CCC to have a sensitivity of .93 and specificity of .46, whereas the SRS-2 sensitivity was .78 and specificity was .67. The authors note, as these findings suggest, that as a screening measure the CCC shows greater emphasis in identifying more children at risk, although with higher rates of false positives. However, they also state that the SRS-2 performs better when used in a clinical setting for diagnostic decision making.

The authors provide several studies comparing the SRS-2 with the ADOS and ADI-R, which uses a behavioral observation design to evaluate for ASD. In a small group of individuals diagnosed with ASD ($n = 61$), correlations between the SRS-2 and ADI-R ranged from .65 to .77 for mother reports, .60 to .74 for father reports, and .52 to .70 for teacher reports (Constantino & Todd, 2003). Lower findings were found in a subsequent larger sample with correlations ranging from .31 to .36 for parent reports and .26 to .40 for teacher reports (Constantino et al., 2007). The manual also provides several studies for European samples comparing the SRS-2 and ADI-R that demonstrated similar results. Comparing the SRS-2 and ADOS domain scores, correlations with parent scores ranged from .37 to .58, whereas teacher-report scores ranged from .15 to .35 (Constantino et al., 2007). Bölte et al. (2011) found a correlation of $r = .48$.

Studies were also presented that compared the SRS-2 with the Child Behavior Checklist (CBCL; Achenbach, 1991) and VABS (Sparrow et al., 1984). Constantino, Przybeck, Friesen, and Todd (2000; $n = 84$) and Bölte et al. (2008; $n = 119$) compared the CBCL with the SRS-2, reporting that the SRS-2 was more sensitive to behaviors associated with ASD, less sensitive to behaviors seldom seen in ASD, and greater sensitivity to behaviors not assessed by the CBCL. Both studies found moderate correlations on the SRS-2 and CBCL subscales ranging from .48 to .64. These correlations largely overlapped with the CBCL subscales Social Problems, Thought Problems, and Attention Problems. As expected, correlations were lower

for those subscales not associated with behaviors seen in those diagnosed with ASD. Studies comparing the SRS-2 and VABS demonstrated a correlation of .44 for the VABS composite score (Charman et al., 2007). A study by Bölte et al. (2008) found a composite correlation of $r = .36$ and subscale correlations ranging from .34 to .43. The manual suggests these correlations with the VABS would be expected, given developmental impairments associated with ASD.

The manual authors conducted a study to evaluate the placement of the 65 SRS-2 items into five treatment subscales: Social Awareness, Social Cognition, Social Communication, Social Motivation, and Restricted Interests and Repetitive Behaviors. Expert judges ($N = 25$), including counselors, social workers, psychiatrists, pediatricians, and psychologists who had experience in working with ASD and PDD, were given the 65 items and asked to sort each item into one of the five groups. Each item was given an expert assignment based on the majority of placements. In order to compare the original placements with the expert placements, nominal scale cross-tabulation was used—there was a significant result ($\chi^2 = 94.24$, $p < .001$). Proportionalreduction-in-error statistics yielded Cohen's kappa of .585 and lambda = .506 (for both, $p < .001$).

The manual also reports that because subscales were not created as fully independent measures, there is a high degree of intercorrelation among them. Parent-report data from 168 cases were used to assess the consistency between the item-to-scale assignments. Alpha reliabilities were calculated for the set of items in each subscale. Values ranged from .60 in the Social Motivation subscale (11 items) to .72 for the Restricted Interests and Repetitive Behaviors subscale (12 items). In addition, the correlations of items with their subscale membership versus other subscales were examined. The authors concluded that there was support for the assignment of items to their respective scales.

Discriminant validity—the extent to which the SRS differentiates ASD from other psychiatric disorders—was assessed in several studies (Kalb, Law, Landa, & Law, 2010; Constantino et al., 2000; Charman et al., 2007; Coon et al., 2010; Bölte et al., 2008, 2011; Kamio et al., 2013; Pine et al., 2008; Towbin, Pradella, Gorrindo, Pine, & Leibenluft, 2005; Reiersen, Constantino, Volk, & Todd, 2007; Reiersen, Constantino, Grimmer, Martin, & Todd, 2008; Puleo & Kendall, 2011; Granader et al., 2010; Hilton et al., 2010; Hilton, Crouch, & Israel, 2008). The manual provides these mean scores and standard deviations, as well as descriptions for each study. The manual also provides a table demonstrating four largely nonoverlapping groups of the SRS-2 mean scores that demonstrate effectiveness in differentiating symptomatology. Groups of normal/typically developing males, females, and combined samples and control groups and normative data found mean total raw scores to range from 14.7 to 34. Those groups

with behavioral and/or emotional diagnoses (e.g., ADHD, anxiety, major depression, mixed diagnoses, mood disorder) without ASD had mean scores that ranged from 40 to 69.2. Studies of those with other ASD, Asperger's, and PPD-NOS demonstrated mean scores from 77.6 to 101.47. Last, studies of those who were autistic provided mean scores from 89.9 to 116.1.

The SRS-2 Adult (Relative/Other-Report) and Adult (Self-Report) forms were evaluated for mean differences in four studies. Lyall (2011) reported controls with no ASD-related diagnoses had mean total scores of .19.4 (*SD* = 18), whereas those with ASD-related diagnoses had mean total scores of 98.8 (*SD* = 34). Those with ASD-related diagnoses were further evaluated for autistic (*M* = 110.5, *SD* = 31), Asperger's disorder (*M* = 93.4, *SD* = 33), and PPD-NOS (*M* = 99.6, *SD* = 35). On the SRS-2 Adult (Relative/Other-Report) form, Bolte (2011) provided information on 240 individuals in three groups: ASD (*n* = 20; *M* = 78.5, *SD* = 13.7), mixed psychiatric (*n* = 62; *M* = 63.4, *SD* = 15.4), and typically developing (*n* = 163; *M* = 55.5, *SD* = 9.9). In a large group (*n* = 250) of adults 18–36 years old diagnosed as autistic, Seltzer et al. (2011) reported a mean score of 94.6 (*SD* = 28.5). Another study (Mandell et al., 2012) comparing a group of individuals diagnosed with ASD (*n* = 14) and a mixed psychiatric sample (primarily schizophrenia; *n* = 127) found mean scores of 100.2 (*SD* = 32.7) and 76.5 (*SD* = 32.5), respectively. Mandell et al. (2012) also conducted a ROC analysis and found a sensitivity of .86 and specificity of .60.

In providing psychometric and validation evidence for the SRS-2 Preschool form, the authors note difficulties in collecting ASD-specific data due to the inherent problems of those who may not display symptoms of ASD until older. As no studies were conducted at the time of manual publication on those with clinical disorders (e.g., ASD, internalizing or externalizing disorders), the authors note that only results from the standardization study can serve as a reference point. One study found convergent validity with the CARS that provided interclass correlations to be .41 ($n = 21$, $p < .002$).

CONCLUSIONS

The information summarized in this chapter provides researchers and clinicians with important characteristics of methods used to assess behaviors associated with ASD, as well as a review of the psychometric qualities that such measures should possess. Table 2.2 provides a summary of the essential aspects of these instruments. As is apparent from examination of the table and the reviews provided earlier in this chapter, the authors of these rating scales differ considerably in their approach to instrument development. For instance, some of the scales are very short (e.g., the CARS-2 has

TABLE 2.2. Comparison of Essential ASD Rating Scale Characteristics

Behavior rating scale	No. of items	Age range	Comparison sample size	Comparison sample	Representative standardization sample	Scores for total scale	Scores for scales
Autism Diagnostic Interview—Revised (ADI-R)	93	2–*x* years	Exact *N* not given	Children with and without ASD, studies conducted by authors where interviews were administered as part of routine initial clinical assessment and systematic research evaluations	No	Raw score	Summary raw scores
Autism Spectrum Rating Scale (ASRS)	80	2–5 and 6–18 years	2,560	National standardization sample of children and youth in the United States and Canada	Yes	*T*-score	*T*-scores
Childhood Autism Rating Scale (CARS)	15	Exact ages not given	1,600	Children who were referred to the TEACCH program (see text)	No	Raw score	None
Social Communication Questionnaire (SCQ)	40	4–*x* years	200	A wide variety of individuals (persons with autism, atypical autism, Asperger syndrome, fragile X syndrome, Rett syndrome, conduct disorder, language delay, intellectual disability, and other clinical diagnoses)	No	Raw score	Raw scores
Social Responsiveness Scale (SRS)	65	4–18 years	1,636	Cases from five studies, combined into one sample (74% white, 11% black, 11% Hispanic, 2% Asian, 2% other)	No	*T*-score	*T*-scores

only 15 items), whereas others contain many items (e.g., 93 items in the ADI-R). Some authors provide only raw scores, which makes interpretation difficult, and only two scales (the ASRS and SRS-2) provide standard scores (*T*-scores). Although these two tests provide derived scores, only the sample upon which the ASRS was based was selected to represent the normal population. Basing standard scores on a national sample is greatly preferable to basing them on a sample of individuals who may have autism.

All the scales except the ASRS and SRS-2 use children with suspected or verified psychological disorders from either research studies or clinic settings as a comparison group. This method allows a clinician to determine whether an examinee is like other children with suspected or documented psychological problems, but comparing the score a child gets on a rating scale to the scores of other children who (1) were referred for evaluation, (2) had some diagnosis on the autism spectrum, or (3) participated in a study of children with autism has several problems. If an individual gets a *T*-score of 50, this would mean that he or she has evidenced behaviors like those of persons who may have ASD. This is not a diagnostic statement, however, for two reasons. First, there is no evidence that the samples used to create the comparison groups for each scale are representative of children with ASD or of the U.S. population. The samples may be limited in demographic characteristics, and therefore the comparison will be affected by the variability of that sample. The sample may be restricted or very heterogeneous, either of which will (1) be undetectable and (2) have a considerable effect on the quality of the comparison. Second, because it is unknown how well such a sample represents children and adolescents with ASD in the particular state in which the sample was collected, or any other state, generalization to clients in other states is limited.

Using a national sample to construct a norm conversion table provides a considerable advantage, for several reasons. First, a large sample allows for reliable calibration of derived scores. Second, comparison to that sample yields an understanding of how often behaviors associated with ASD are found within the typical population. Third, the comparison of a child's or adolescent's behavior to what is expected in the typically developing population provides for greater understanding of how far an individual may be from the norm. Fourth, having a wellnormed score provides a means of calibrating how much response to intervention is needed to bring the individual's behavior into a range that can be considered typical.

The most glaring shortcoming of nearly all these scales is that they do not have standard scores that are based on a national standardization sample. This possesses a considerable liability for those who choose to use these measures, because it is imperative to know how different an examinee's behavior is from that of typical individuals, as well as how the behaviors compare with those of persons with ASD. The only way to know the rate

at which typical children show behaviors associated with ASD is to have a national standardization group and to base norms on this sample. Clinicians can then make defensible statements about how far a child deviates from normality and to what extent the normative data support a diagnosis. Those measures that do not have a national standardization sample should be viewed with caution by clinicians, because interpretation of results across tests is made very difficult by the differences in the samples, and the stability of the norms cannot be determined. The use of welldeveloped, psychometrically sound assessments will greatly enhance the likelihood that accurate and valid information can be obtained.

REFERENCES

Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist/4-18 and 1991 profile*. Burlington: University of Vermont, Department of Psychiatry.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing, 2014 edition*. Washington, DC: American Educational Research Association.

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.

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

Berument, S. K., Rutter, J., Lord, C., Pickles, A., & Bailey, A. (1999). Autism screening questionnaire: Diagnostic validity. *British Journal of Psychiatry, 175,* 444–451.

Bishop, D. V. M., & Norbury, C. F. (2002). Exploring the borderlands of autistic disorder and specific language impairment: A study using standardized diagnostic instruments. *Journal of Child Psychology and Psychiatry, 43,* 917–929.

Bölte, S., Poustka, F., & Constantino, J. N. (2008). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Bölte, S., Westerwald, E., Holtmann, M., Freitag, C., & Poustka, F. (2011). Autistic traits and autism spectrum disorders: The clinical validity of two measures presuming a continuum of social communication skills. *Journal of Autism and Developmental Disorders, 41*(1), 66–72.

Bracken, B. A. (1987). Limitations of preschool instruments and standards for minimal levels of technical adequacy. *Journal of Psychoeducational Assessment, 5,* 313–326.

Bracken, B. A., & McCallum, R. S. (1997). *Universal Nonverbal Intelligence Test*. Itasca, IL: Riverside.

Charman, T., Baird, G., Simonoff, E., Loucas, T., Chandler, S., Meldrum, D., et al. (2007). Efficacy of three screening instruments in the identification of autistic-spectrum disorders. *British Journal of Psychiatry, 191*(6), 554–559.

Constantino, J. N., Davis, S. A., Todd, R. D., Schindler, M. K., Gross, M. M., Brophy, S. L., et al. (2003). Validation of a brief quantitative genetic measure of autistic traits: Comparison of the Social Responsiveness Scale with the Autism Diagnostic Interview—Revised. *Journal of Autism and Developmental Disorders, 33,* 427–433.

Constantino, J. N., & Gruber, C. P. (2012). *Social Responsiveness Scale, Second Edition manual*. Torrance, CA: Western Psychological Services.

Constantino, J. N., LaVesser, P. D., Zhang, Y., Abbacchi, A. M., Gray, T., & Todd, R. D. (2007). Rapid quantitative assessment of autistic social impairment by classroom teachers. *Journal of the American Academy of Child and Adolescent Psychiatry, 46*(12), 1668–1676.

Constantino, J. N., Przybeck, R., Friesen, D., & Todd, R. D. (2000). Reciprocal social behavior in children with and without pervasive developmental disorders. *Journal of Developmental and Behavior Pediatrics, 21,* 2–11.

Constantino, J. N., & Todd, R. D. (2003). The genetic structure of reciprocal social behavior. *American Journal of Psychiatry, 157,* 2043–2045.

Coon, H., Villalobos, M. E., Robison, R. J., Camp, N. J., Cannon, D. S., Allen-Brady, K., et al. (2010). Genome-wide linkage using the Social Responsiveness Scale in Utah autism pedigrees. *Molecular Autism, 1*(1), 8.

Creek, M. (1961). Schizophrenia syndrome in childhood: Progress report of a working party. *Cerebral Palsy Bulletin, 3,* 501–504.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart & Winston.

Gilliam, J. (2001). *Gilliam Asperger's Disorder Scale manual*. Austin, TX: PRO-ED.

Gilliam, J. (2006). *GARS-2: Gilliam Autism Rating Scale* (2nd ed.). Austin, TX: PRO-ED

Goldstein, S., & Naglieri, J. A. (2009). *Autism Spectrum Rating Scale*. Toronto, ON, Canada: Multi-Health Systems.

Granader, Y. E., Bender, H. A., Zemon, V., Rathi, S., Nass, R., & MacAllister, W. S. (2010). The clinical utility of the Social Responsiveness Scale and Social Communication Questionnaire in tuberous sclerosis complex. *Epilepsy and Behavior, 18*(3), 262–266.

Hilton, C. L., Crouch, M. C., & Israel, H. (2008). Out-of-school participation patterns in children with high-functioning autism spectrum disorders. *American Journal of Occupational Therapy, 62*(5), 554–563.

Hilton, C. L., Harper, J. D., Kueker, R. H., Lang, A. R., Abbacchi, A. M., Todorov, A., et al. (2010). Sensory responsiveness as a predictor of social severity in children with high functioning autism spectrum disorders. *Journal of Autism and Developmental Disorders, 40*(8), 937–945.

Kalb, L. G., Law, J. K., Landa, R., & Law, P. A. (2010). Onset patterns prior to 36 months in autism spectrum disorders. *Journal of Autism and Developmental Disorders, 40*(11), 1389–1402.

Kamio, Y., Inada, N., Moriwaki, A., Kuroda, M., Koyama, T., Tsujii, H., et al. (2013). Quantitative autistic traits ascertained in a national survey of 22,529 Japanese schoolchildren. *Acta Psychiatrica Scandinavica, 128*(1), 45–53.

Kanner, L. (1943). Autistic disturbances of affective contact. *Nervous Child, 2,* 217–250.

Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Assessment Battery for Children, Second Edition manual*. Circle Pines, MN: American Guidance Service.

Krug, D. A., Arick, J. R., & Almond, P. J. (2008). *Autism Behavior Checklist, Second Edition*. Austin, TX: 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*. Torrance, CA: Western Psychological Services.

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

Mandell, D. S., Lawer, L. J., Branch, K., Brodkin, E. S., Healey, K., Witalec, R., et al. (2012). Prevalence and correlates of autism in a state psychiatric hospital. *Autism, 16*(6), 557–567.

Naglieri, J. A., Das, J. P., & Goldstein, S. (2014). *Cognitive Assessment System, Second Edition*. Itasca, IL: Riverside.

Naglieri, J. A., Das, J. P., & Goldstein, S. (2013). *Cognitive Assessment System* (2nd ed.). Austin, TX: PRO-ED.

National Society for Autistic Children. (1978). National Society for Autistic Children definition of the syndrome of autism. *Journal of Autism and Developmental Disorders, 8*, 132–137.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.

Pilowsky, T., Yirmia, N., Shulman, C., & Dover, R. (1998). The Autism Diagnostic Interview—Revised and the Childhood Autism Rating Scale: Differences between diagnostic systems and comparison between genders. *Journal of Autism and Developmental Disorders, 28*(2), 143–151.

Pine, D. S., Guyer, A. E., Goldwin, M., Towbin, K. A., & Leibenluft, E. (2008). Autism spectrum disorder scale scores in pediatric mood and anxiety disorders. *Journal of the American Academy of Child and Adolescent Psychiatry, 47*(6), 652–661.

Puleo, C. M., & Kendall, P. C. (2011). Anxiety disorders in typically developing youth: Autism spectrum symptoms as a predictor of cognitive-behavioral treatment. *Journal of Autism and Developmental Disorders, 41*(3), 275–286.

Reiersen, A. M., Constantino, J. N., Grimmer, M., Martin, N. G., & Todd, R. D. (2008). Evidence for shared genetic influences on self-reported ADHD and autistic symptoms in young adult Australian twins. *Twin Research and Human Genetics, 11*(6), 579.

Reiersen, A. M., Constantino, J. N., Volk, H. E., & Todd, R. D. (2007). Autistic traits in a population-based ADHD twin sample. *Journal of Child Psychology and Psychiatry, 48*(5), 464–472.

Ritvo, R. A., Ritvo, E. R., Guthrie, D., Rito, M. J., Hufnagel, D. Y., McMahon, W., et al. (2011). The Ritvo Autism Asperger Diagnostic Scale—Revised (RAADS-R): A scale to assist the diagnosis of autism spectrum disorder in adults: An international validation study. *Journal of Autism Developmental Disorders, 41*(8), 1076–1089.

Rutter, M. (1978). Diagnosis and definition of childhood autism. *Journal of Autism and Developmental Disorders, 8*, 139–161.

Rutter, M., Bailey, A., & Lord, C. (2003). *Social Communication Questionnaire*. Los Angeles: Western Psychological Services.

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

Schopler, E., Lansing, M. D., Reichler, R. J., & Marcus, L. M. (2005). *Psychoeducational Profile, Third Edition*. Austin, TX: PRO-ED.

Schopler, E., Van Bourgondien, M. E., Wellman, G. J., & Love, S. R. (2010). *Childhood Autism Rating Scale, Second Edition*. Torrance, CA: Western Psychological Services.

Seltzer, M. M., Greenberg, J. S., Taylor, J. L., Smith, L., Orsmond, G. I., Esbensen, A., et al. (2011). Adolescents and adults with autism spectrum disorder. In D.

G. Amaral, G. Dawson, & G. Geschwind (Eds.), *Autism spectrum disorders* (pp. 242–252). New York: Oxford University Press.

Sparrow, S. S., Balla, D. A., & Cicchetti, D. V. (1984). *Vineland Adaptive Behavior Scales*. Circle Pines, MN: American Guidance Services.

Thorndike, R. L. (1982). *Applied psychometrics*. Boston: Houghton Mifflin.

Towbin, K. E., Pradella, A., Gorrindo, T., Pine, D. S., & Leibenluft, E. (2005). Autism spectrum traits in children with mood and anxiety disorders. *Journal of Child and Adolescent Psychopharmacology, 15*(3), 452–464.

U.S. Census Bureau. (2000). *Statically abstract of the United States: 2000*. Washington, DC: U.S. Government Printing Office.

U.S. Census Bureau. (2009). *Statically abstract of the United States: 2009*. Washington, DC: U.S. Government Printing Office.

Wechsler, D. (2014). *Wechsler Intelligence Scale for Children, Fourth Edition*. San Antonio, TX: Psychological Corporation.

World Health Organization. (1992). *The ICD-10 classification of mental and behavioral disorders: Clinical descriptions and guidelines*. Geneva, Switzerland: Author.