CHAPTER THREE

Psychometric Issues and Current Scales for Assessing Autism Spectrum Disorder

Jack A. Naglieri
Kimberly M. Chambers

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 disorders (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 understand ASD bet-

ter. 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 to 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 95 (100 – 5) to 105 (100 + 5). Figure 3.1 provides confidence intervals (95% level of confidence) for a standard score of 100 that would be obtained for measures with reliability of .50 through .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—Fourth Edition (Wechsler, 2003) and the Cognitive Assessment System (Naglieri & Das, 1997—the confidence intervals are provided in the test manual's table for converting sums of subtest scores

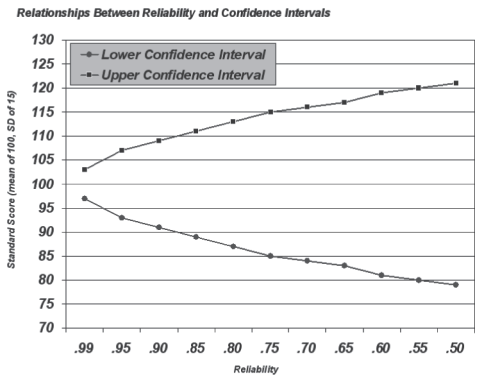

FIGURE 3.1. Relationships between reliability and confidence intervals.

to standard scores, and the range is already centered on the estimated true score. The relationships among the various scores are illustrated in Table 3.1, which provides the obtained score, estimated true score, and lower and upper ranges of the 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 3.1, the range for a standard score of 80 is 74 to 90 (6 points below 80 and 10 points above 80). In contrast scores for standard scores that are above the mean are *lower* than the obtained score. The range for a standard score of 120 is 110 to 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.

TABLE 3.1. 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.

The importance of the *SEM* becomes most relevant when two scores are compared. The lower the reliability, the larger the *SEM*, and the more likely an individual's scores are to differ on the basis of chance. For example, when a child's score on a measure of self-regulation is compared to 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 self-regulation (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 3.2 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 coef-

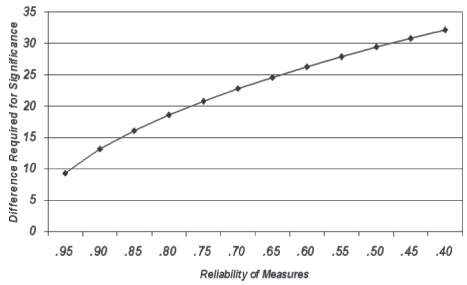

FIGURE 3.2. Relationships between reliability and the differences needed for significance when one is comparing two scores. Note that this figure assumes two variables with the same reliability and an *SD* of 15 at the 95% level of confidence.

ficient 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 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 long-lasting 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,* fourth edition, text revision [DSM-IV-TR]; American Psychiatric Association, 2000) will undoubtedly evolve to reflect future research findings. Similarly, research methodology is also important, particularly when typical children are being compared to 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-IV-TR). 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 [AERA], American Psychological Association, & National Council on Measurement Education, 1999), evidence for validity "integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses" (p. 17). There are 24 standards relating to validity issues that should 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 (Naglieri & Das, 1997). 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 (AERA et al., 1999). 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 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 type 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 of this chapter, 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 the chapter with a commentary on the relative advantages of these scales.

Descriptions of Scales Used to Assess ASD

Autism Diagnostic Observation Schedule

Description

The Autism Diagnostic Observation Schedule (ADOS; Lord, Rutter, DiLavore, & Risi, 2002) is a semistructured assessment of communication, social interaction, and play in children or adults suspected of having ASD. The present ADOS, the ADOS-Generic or ADOS-G, is a combination of the 1989 ADOS (Lord et al., 1989) and the Pre-Linguistic Autism Diag-

nostic Observation Schedule (PL-ADOS; DiLavore, Lord, & Rutter, 1995). A referred individual is assessed with one of the four modules contained in the ADOS. Each module can be administered in 30–45 minutes and is geared for a child or adult at a particular developmental and 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.

Module 1 is most appropriate for children who are at the preverbal or single-word language level. It 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 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 5 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 adolescents and adults.

Examiners take notes during ADOS 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. The ADOS uses cutoff scores for two separate domains (Social Interaction and Communication), as well as a Communication–Social Interaction total cutoff score, in order to make the diagnostic distinction between autism and the broader category of ASD. Although the ADOS has many similarities to the DSM-IV-TR and *International Classification of Diseases,* 10th revision (ICD-10) models of diagnosing ASD, the ADOS algorithm included in the manual does not include a measure of restricted, repetitive, and stereotyped patterns of behavior identified by the DSM-IV-TR and ICD-10. However, these behaviors are coded in a separate domain called Stereotyped Behaviors and Restricted Interests. In addition, the ADOS does not include information about the age of onset or early history required for a DSM-IV-TR or ICD-10 diagnosis. Recently, however, new algorithms for the ADOS have been published (Gotham, Risi, Pickles, & Lord, 2007) that do include restricted/repetitive/stereotyped behaviors and no longer have separate Social Interaction and Communication domains.

Description of the Comparison Group

The validation sample of the ADOS consisted of individuals who were referred to the Developmental Disorders Clinic at the University of Chicago. These individuals were evaluated and assigned a clinical diagnosis

of autism, pervasive developmental disorder not otherwise specified (PDD-NOS), or "non-spectrum" by a child psychologist and child psychiatrist through various measures and observations. In addition to the initial sample of participants, individuals were also recruited from various other locations in order to obtain three samples (autism, PDD-NOS, and non-spectrum) in each module that were of adequate size and roughly equivalent in verbal mental age or verbal IQ. Within each module, participants were chosen from the three groups to constitute samples that would be similar with regard to chronological age, gender, and ethnicity. Participants in the validity study were then selected for inclusion in one of four module samples on the basis of verbal ability. In contrast to the other modules, Module 1 included a group with low-functioning autism because of the importance of documenting autism in very young, severely delayed children.

The composition of the sample in each module was as follows:

Module 1: lower functioning autism (n = 20), matched autism (n = 20), PDD-NOS (n = 17), and non-spectrum (n = 17)
Module 2: autism (n = 21), PDD-NOS (n = 18), and non-spectrum (n = 16)
Module 3: autism (n = 21), PDD-NOS (n = 20), and non-spectrum (n = 18)
Module 4: autism (n = 16), PDD-NOS (n = 14), and non-spectrum (n = 15)

For each 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 in the study were comparable across modules and groups and were as follows: European American (80%), African American (11%), Hispanic (4%), Asian American (2%), and other or mixed ethnic groups (2%). All participants in the study were native English speakers. In addition, all participants had nothing more than mild hearing or visual impairments and were all ambulatory. Finally, with the exception of one boy with Williams syndrome, there were no participants with identifiable syndromes.

Reliability

In order to evaluate the reliability of individual items on the ADOS, the test authors obtained interrater reliability information for each module. 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 (mean = .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, kappa for 15 items exceeded .60 (mean = .70), and kappa for the remaining items equaled .50, with the exception of 4 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 (mean = .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.

Finally, Module 4 had a minimum of 80% exact agreement, with kappa coefficients exceeding .60 for 22 of the items (mean = .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 were computed for algorithm subtotals and totals for each module and the combined modules. For the separate modules, interclass correlations ranged from .88 to.97 for the Social Interaction domain, .74 to .90 for the Communication domain, .84 to .98 for the Communication–Social Interaction total, and .75 to .90 for Stereotyped Behaviors and Restricted Interests.

Interrater agreement for diagnostic classification for autism versus non-spectrum was examined. For Modules 1 and 3, agreement was 100%; for Module 2, agreement was 91%; and for Module 4, agreement was 90%. However, when participants with PDD-NOS were included, agreement dropped: It was 93% for Module 1, 87% for Module 2, 81% for Module 3, and 84% for Module 4. The authors reported that when Fisher's exact test was used to compare the diagnostic groups, results were significant at $p < .01$, and that disagreements were mostly between PDD-NOS and autism.

When data were collapsed across all modules, interrater correlations for domain and total scores ranged from .82 to .93. In addition, interclass correlations of ratings during the testing session with ratings immediately after testing ranged from .80 to .92. These ratings were made by observing videotapes of the same administration. The interclass correlations ranged from .72 to .92. Finally, test–retest interclass correlations ranged from .59 to .82. Test–retest periods averaged approximately 9 months. The mean differences in domain scores for time 1 and time 2 were 1.19 (*SD* = 1.6) for Communication, 1.26 (*SD* = 1.39) for Stereotyped Behaviors and Restricted Interests, 1.78 (*SD* = 1.93) for Social Interaction, and 2.67 (*SD* = 1.93) for Communication–Social Interaction total. Group means changed less than 0.50 for each domain, with the exception of the Communication–Social Interaction total (*M* = –0.94, *SD* = 2.63). Scores for 6 children in the test–retest sample changed ADOS classification.

Validity

The authors of the ADOS 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. Few details are provided in the manual.

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 ranged from a low of 14 to a high of 21. These findings were augmented by analyses of classification rates. The sample sizes for these analyses by module for the groups based on clinical diagnosis (lower-functioning autism, autism, PDD-NOS, non-spectrum) ranged from 0 to 21. The results of these analyses, which are provided for various combinations and cutoff scores for the domains measured by the ADOS, generally suggest that the instrument had specificity values in the upper 80% to low 90% range, and sensitivity in the upper 90% range.

Autism Diagnostic Interview—Revised

Description

The Autism Diagnostic Interview—Revised (ADI-R; Rutter, Le Couteur, & Lord, 2003b) 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 focuses 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 Interaction; Qualitative Abnormalities in Communication; Restrictive, Repetitive, and Stereotypic 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 Patterns of Behavior 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, 2003a; 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 the full DSM-IV/ICD-10 criteria for autism (this applied to 8 out of a total sample of 21 children and 8 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 8 children meeting the full criteria on the ADI-R, 6 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 Rating Scale

Description

The Autism Rating Scale (ARS; Goldstein & Naglieri, 2008) is an observer-completed rating scale designed to aid in the diagnosis of individuals who may have ASD. The ARS is completed by parents (or similar caregivers) or teachers (or similar professionals) who rate behaviors characteristic of children ages 2–6 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 (Self-Regulation, Social/Communication, and Stereotypical Behaviors) and an ARS 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 *SD* of 10. In addition, a short Screening version of the ARS is provided, consisting of 15 items.

The authors state that the ARS was developed to measure ASD and autism-related problems, in order to allow clinicians to compare an individual to norm-based expectations in an objective and reliable manner. Because the ARS 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 ARS 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 ARS Screening scale is intended to be used in large-scale prevention programs.

Description of the Comparison Group

The ARS 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.

Reliability

The internal reliability coefficients for the empirically based scales for the Early Childhood ARS are as follows: Stereotypical Behaviors (25 items; reliability of .79 for parents and .77 for teachers), Social/Communication (21 items; reliability of .83 and .85 for parents and teachers, respectively), and Self-Regulation (11 items; reliability of .75 for parents and .80 for teachers). The internal reliability coefficients for the empirically based scales for the School Age ARS are as follows: Stereotypical Behaviors (19 items; reliability of .81 for parents and .83 for teachers), Social/Communi-

cation (21 items; reliability of .86 and .88 for parents and teachers, respectively), and Self-Regulation (14 items; reliability of .79 for parents and .82 for teachers). The 15-item ARS 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

At the time of this writing the ARS is in final stages of development, and the validity studies have yet to be completed. Readers interested in seeing the results of the item factor analysis, comparison of the scores obtained on the ARS with other measures of ASD, rates of accurate differentiation of children with and without ASD using the ARS, and other validity studies should see the manual.

Childhood Autism Rating Scale

Description

The Childhood Autism Rating Scale (CARS; Schopler, Reichler, & Renner, 1988) is a 15-item behavior rating scale developed to help identify children with autism and to evaluate varying degrees of the disorder. The CARS was also developed to differentiate autistic children from those with other developmental disorders, particularly those with moderate to severe mental retardation. CARS 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 are summed and categorized as follows: 15–29.5 is considered the nonautistic range, 30–36.5 is considered the range of mild to moderate autism, and 37–60 is considered the severely autistic range. The 15 items included in the CARS are based on the diagnostic criteria from Kanner (1943), the nine dimensions by Creek (1961), Rutter's (1978) definition, and the criteria proposed by the National Society for Autistic Children (1978).

Description of the Comparison Group

The CARS scores are based on a comparison to the ratings of over 1,500 children who were referred to the North Carolina program for the Treatment and Education of Autistic and related Communication-handicapped

Children (TEACCH). 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 24.3% females and 75.7% males, whose racial background was European American (66.9%), African American (30.2%), or other (2.9%). The authors describe the sample as predominantly from low socioeconomic levels, based on the Hollingshead–Redlich two-factor index. Over one-fourth (26.3%) of the sample fell in the lowest socioeconomic category (V) identified by the index. The rest of the sample was distributed as follows: IV (33%), III (22.4%), II (9.3%), and I (9.1%). The sample was further described on the basis of IQ as follows: 70.6% < 70, 16.5% = 70–84, and 12.8% = 85 and above. Finally, the children varied in age as follows: < 6 years = 56.4%, 6–10 years = 32.0%, and 11 and above = 11.4%. 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 manual.

Reliability

The CARS manual presents internal, test–retest, and interrater reliability coefficients. Internal reliability is reported, but the manual does not specify for what sample the coefficient (alpha = .94) was calculated. Test–retest reliability was computed for 91 individuals assessed over a period of about 1 year, but details of the exact test–retest interval and characteristics of the sample are not provided. There were no significant differences between the mean raw score earned at each rating, and the reliability coefficient (kappa) was .64. The average interrater reliability was .71 for a sample of 280 individuals (no further information on the sample is provided in the test manual). Interrater reliabilities for each of the 15 items ranged from .62 to .93.

Validity

The authors assessed criterion-related validity for the 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 validity of CARS ratings made under alternate conditions was also examined. Because the CARS was originally developed to be used during the administration of the Psychoeducational Profile (PEP), different groups of children were rated on the basis of information gathered both during a PEP session and in a parent interview, a classroom observation, or a chart history. Children ($N = 41$) were rated by a therapist after a meeting with each child's parents. CARS scores that were based on the parent interview were compared to CARS scores during the PEP session. There was no significant difference between the two scores (PEP mean = 32.7; interview mean = 33.7; $t = -1.26$, $p > .10$). Correlation between the scores indicated good agreement ($r = .82$, $p < .01$). CARS screening diagnosis from the parent interview and PEP administration agreed in 90% of the cases (kappa coefficient was .75).

Raters visited classrooms for 1- to 2-hour observations of 20 children who would also receive the PEP in the clinic. Mean ratings based on observations in the classroom did not differ significantly from mean ratings based on observations made during PEP administration (PEP $\overline{X} = 32.48$, classroom $\overline{X} = 34.18$; $t = -1.55$, $p > .10$). The correlation of the ratings was .73 ($p < .01$). The classroom observations and PEP administration agreed in 86% of the cases (kappa coefficient was .86).

Raters also provided CARS ratings using the behavioral information contained in the case history charts of 61 children and by PEP administration. The mean ratings did not differ significantly (PEP $\overline{X} = 32.34$, chart review $\overline{X} = 32.47$; $t = 0.20$, $p > .10$). The correlation of these ratings was .82 ($p < .01$). CARS screening diagnosis using the two methods agreed in 82% of the cases (kappa coefficient was .63).

The validity of CARS ratings made by professionals in other disciplines was also examined. Professionals in related fields were given brief introductions to the CARS and asked to make ratings based on their observations of behavior during PEP administration. One hour prior to observations, professionals read the CARS manual and observed a training video. Ratings made by visiting professionals were compared with the criterion ratings made by clinical directors observing the same session. The 18 visiting professionals consisted of medical students, pediatric residents and interns, special educators, school psychologists, speech pathologists, and audiologists. The mean CARS ratings scores were not significantly different from the mean of the clinical directors (visitor $\overline{X} = 32.46$, clinical director $\overline{X} = 33.15$; $t = 0.92$, $p > .10$). The scores were also correlated with each other ($r = .83$, $p < .01$). Diagnostic screening categorizations resulting from CARS ratings of the two groups showed 92% agreement (kappa coefficient was .81).

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 (6 measuring developmental abilities and 4 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 autistic children. 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 an 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 148 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 the ages of 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 148 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 through 11). Average alpha coefficients for Performance 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 autistic children between the ages of 4 and 14 residing in California, Oklahoma, and Texas. The sample consisted of 28 males and 5 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 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 through 10 from seven different states. Of the 40 participants, 33 were male and 7 were female; 1 was Hispanic, 32 were white, 6 were black, and 2 were of other races. Nine of the 40 children did not have a disability, 29 were diagnosed with autism, and 2 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 (mean = .85); Personal Self-Care items correlations ranged from .65 to 1.00 (mean = .90); and correlations for the Adaptive Behavior subtest items ranged from .52 to .90 (mean = .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 through 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 nonblack, and Hispanic versus non-Hispanic. 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 (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 et al., 2003a) 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 at al., 2003b), 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 three 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 an 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; Restricted, Repetitive, and Stereotyped Behaviors and Interests). Although the produc-

tion 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 to 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, 7 had fragile X syndrome, 5 had Rett syndrome, 10 had conduct disorder, 7 had language delay, 15 had mental retardation, and 8 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 5 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 statistically different 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 (self-injury and unusual attachment to objects) differentiated at the 7% significance level and showed more modest correlations with the total score (.37 and .27). 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 et al., 1999)

Three- and four-factor solutions were explored for 39 items of the SCQ (items 2–40). Analysis suggested that a four-factor structure appeared to be an acceptable fit. Principal-component 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, 8.7% (eigenvalue = 3.38) by a communication factor, 5% (eigenvalue = 1.94) by an abnormal language factor, and 4.5% (eigenvalue = 1.74) by a stereotyped behavior factor. 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 to .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).

Receives operating characteristics (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 mental retardation (AUC = .86). The SCQ also effectively differentiated between autism and non-ASD conditions other than mental retardation (AUC = .94), autism and mental retardation (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 to that of the ADI-R. AUC results were contrasted for ASD versus non-ASD conditions (AUC = .88 and .87, respectively), autism versus mental retardation (AUC = .93 and .96), and autism versus other ASD (AUC = .73 and .74).

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 the highest in the group with severe mental retardation (14.74), and that 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 mental retardation (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 mental retardation, and a sensitivity of .96 and a specificity of .67 for autism versus mental retardation. 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, and 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 codes 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

Description

The Social Responsiveness Scale (SRS; Constantino & Gruber, 2005) is a 65-item questionnaire that covers various dimensions of interpersonal behavior, communication, and repetitive/stereotypical behavior. The SRS can be used with children ages 4–18 as both a screening tool and an aid

to clinical diagnosis; it helps to identify autistic disorder, Asperger disorder, PDD-NOS, and schizoid personality disorder. The SRS is completed by someone familiar with the child's current behavior and developmental history. Items on the scale focus on the behavior of the child, and the rater's responses are given in a Likert format. There are separate parent and teacher forms, each of which takes approximately 15 minutes to complete and 5–10 minutes to score.

The test authors recommend that the SRS 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 for males and a raw score of 65 for females. When the instrument is used for screening children suspected of having social development problems, a cutoff score of 85 is recommended for both genders. When SRS scores are converted to *T* scores, a value of 60–70 is considered to be in the mild to moderate range and is typical of children with mild or high-functioning ASD. A *T* score of 76 or higher is in the severe range and indicative of a diagnosis of autism. In addition to total *T* scores, the SRS also yields *T* scores for five treatment subscales: Social Awareness, Social Cognition, Social Communication, Social Motivation, and Autistic Mannerisms. Instructions for interpretation of these scores are provided in the test manual.

Description of the Comparison Group

Subjects from five research studies were combined to produce a sample of over 1,600 children to norm the SRS. The first three groups consisted of random samples of twins identified from Missouri birth records. The fourth group was based on parent report data on 145 males and 127 females from elementary, middle, and high schools located in an economically diverse U.S. Midwestern suburban school district. The fifth sample consisted of teacher report data from 552 students taken from a large suburban school district in the Midwest and a large urban district in the West. The racial composition of this normative group was as follows: White (74%), black (11%), Hispanic (11%), Asian (2%), and other (2%). Separate norms based on these data were created for both parent and teacher versions of the SRS.

Reliability

Internal consistency, construct temporal stability, and interrater agreement data are provided by the SRS manual as assessments of reliability. Internal consistency alpha coefficients are provided for various samples of normative parent ratings, normative teacher ratings, and clinical ratings. Parent

and teacher ratings were broken down by gender. The alpha coefficient for parent data for males was .94 (n = 512) and for females was .93 (n = 569). In turn, the alpha coefficient for teacher data for the male population was .97 (n = 278) and for females was .96 (n = 277). Alpha coefficients were not broken down by gender for clinical ratings, and the coefficient for a sample of 281 was .97. Alpha coefficients were also broken down by gender when construct temporal stability was reported. The coefficient for males was .85 (n = 102) and for females was .77 (n = 277). Finally, interrater reliability coefficients were provided from a sample of 63 children. Coefficients were provided for mother–father rater pairings (retest r = .91), mother–teacher pairings (retest r = .82), and father–teacher pairings (retest r = .75).

Validity

The test authors used three independent samples to assess the factor structure of the SRS. These samples included a teacher report from a normal school sample (Constantino, Przybeck, Friesen, & Todd, 2000), an epidemiological sample of male twins (Constantino, Davis et al., 2003), and a clinical sample involving 226 child psychiatric patients with and without pervasive developmental disorders (PDD) (Constantino et al., 2004). Results of the factor analysis of these samples failed to support the existence of independent subdomains of dysfunction in ASD. The analysis supported one factor, which resulted in disparate phenotypic manifestations across the three criterion domains for autistic disorder: social deficits, language deficits, and repetitive/stereotypical behaviors. Over 25 SRS items had factor loadings greater than .60 on this primary factor, including items representing all three DSM-IV criterion domains for autism (social deficits, language deficits, and restricted interests/stereotypical behavior).

The authors also report that ROC analysis revealed high degrees of sensitivity and specificity for both the screening and clinical cutoff scores. Although research on the SRS has identified autistic symptoms that may be attributable to a singular underlying deficiency, SRS treatment subscales were developed to aid in identifying therapeutic needs and approaches. The authors also report that a total raw score of 75 was associated with a sensitivity of .85 and specificity of .75 for any ASD as rated by expert clinicians, and a total raw score of 85 was associated with sensitivity of .70 and a specificity of .90.

The authors also describe a study conducted to evaluate the placement of the 65 SRS items into five treatment subscales: Social Awareness, Social Cognition, Social Communication, Social Motivation, and Autistic Mannerisms. 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 ($\chi^2 = 94.24$, $p < .001$). Proportional-reduction-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 (Constantino et al., 2004). Alpha reliabilities were calculated for the set of items in each subscale. Values ranged from .77 in the Social Awareness subscale (8 items) to .92 for the Social Communication subscale (22 items). In addition, the correlations of items with their subscale membership versus other subscales were examined. The authors concluded that there was some support for the assignment of items to their respective scales.

The structure of the SRS was further examined in a sample of parents ($N = 1,576$) of randomly ascertained twins. The mean scores for males in this sample was 35.3 and for females was 27.5 ($t = 7.63$, $p < .001$). When the previously established cutoff score for males with PDD-NOS was used, 1.4% of males and 0.3% of females had scores at or above the cutoff (Fisher exact test, $p = .03$)(Constantino & Todd, 2003). In addition, Chakrabarti and Fombonne (2001) found the prevalence of PDD to be 0.6%, with a male-to-female ratio of 4:1.

The authors used structural equation modeling to examine the factor structure of the SRS for 232 monozygotic and dizygotic male twins. Intraclass correlations of twin–twin pairs for scores on the SRS were .73 for monozygotic twins and .37 for dizygotic twins. They reported that genetic factors accounted for approximately 76% of the total variance, and that SRS scores were not significantly influenced by age, rater bias, or rater contrast effects (Constantino & Todd, 2000). The authors also reported that the additive genetic influence on social deficits in the epidemiological sample of females was .40. In order to determine whether gender-specific genetic effects accounted for the discrepancy between males and females, the SRS was administered to the parents of 300 of opposite-sex dizygotic twins. The authors reported that genes influencing autistic traits appear to be the same for males and females (Constantino & Todd, 2003).

Data for the Child Behavior Checklist (CBCL) and SRS were also obtained for male twins ($N = 219$). Regression analysis indicated that scores for internalizing and externalizing behavior explained less than 5% of the variance in SRS scores. *SEM* was used to determine that the best-fitting model was one in which the majority of causal influences on SRS scores (55%; 90% confidence interval = .45–.70) were genetic influences specific

to the SRS. Less than 20% of the overlap in the causal influence scores on either instrument was accounted for by overlap in phenotypic characteristics of the SRS and syndromal CBCL scores for attention problems and social problems (Constantino, Hudziak, & Todd, 2003). The authors suggested that the SRS measures a unique, genetically determined component of psychopathology independent from other domains of child psychopathology, and that subthreshold autistic differences may operate to make other psychopathology worse (Constantino & Todd, 2003). The authors further suggest that measuring subthreshold autistic differences can be helpful for predicting a clinical course and understanding influences on other child mental health problems.

In order to further assess the structural validity of the SRS, a preliminary study of the extent to which subthreshold autistic traits measured were related to familial predisposition to autism in the clinical sample (Constantino et al., 2004) was conducted. Parent report data were collected from 72 siblings of the initial clinical sample (48 siblings of participants with PDD, 24 siblings of participants with non-PDD psychiatric diagnoses). The mean SRS score for siblings with PDD (45.9) was substantially higher than that for siblings with non-PDD psychiatric diagnoses (16.8) ($t = 4.09$, $p = .0001$). According to the authors, these results indicate that autistic deficits measurable by the SRS aggregate in the siblings of patients with PDD.

Discriminant validity—the extent to which the SRS differentiates ASD from other psychiatric disorders—was assessed in psychiatric patients ($N = 158$) with and without ASD, as well as 287 randomly selected children from a St. Louis school district who participated in the initial study of the SRS (Constantino et al., 2000). High scores on the SRS were associated with the clinical diagnoses of autistic disorder, Asperger's disorder, and PDD-NOS, and not with other psychiatric conditions or with low IQ. Average SRS scores for children with various noncomorbid Axis I mood disorders were as follows: mood disorder, 59.4; conduct disorder, 48.4; psychotic disorder, 40.3; attention-deficit/hyperactivity disorder, 51.1; and PDD-NOS, 101.5. Children with PDD-NOS had significantly higher scores than children in other groups (single-factor analysis of variance [ANOVA]: $F = 11.69$, $df = 4.75$, $p < .001$).

The SRS manual also describes a study (Constantino, Davis, et al., 2003) to assess the relationships between the SRS and the ADI-R in a clinical sample of 61 child psychiatric patients. Subjects were evaluated with the SRS and the ADI-R a month apart. Participants whose SRS scores score fell within 2 *SD* of the mean (100) for PDD-NOS had ADI-R social deficit scores ranging from 0 to 30 (the ADI-R clinical cutoff being 10). No respondent with an ADI-R score above the clinical cutoff had an SRS score below 65, indicating the presence of symptoms that were at least mildly to moderately clinically significant. Mean scores for clinical participants without

PDD were also lower than scores for participants with PDD (single-factor ANOVA: F =72.95; df = 2.58; $p < .0001$). Significant correlations between mothers, fathers, and teachers on quantitative assessments of autistic deficits with the SRS were found and ranged from .75 to .91. SRS scores were also found to be unrelated to IQ and exhibited a 2-year test–retest stability of .83.

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 3.2 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 example, some of the scales are very short (e.g., the CARS has only 15 items), whereas others contain many items (e.g., 93) for the ADI-R. Some authors provide only raw scores, which make interpretation difficult, and only two scales (the ARS and SRS) provide standard scores (T scores). Although these two tests provide derived scores, the samples upon which they are based reflect a fundamental difference in test development. 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 ARS and SRS 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 autistic children has several problems. First, 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

TABLE 3.2. 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 Rating Scale (ARS)	80	2–5 and 6–18 years	2,000	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, mental retardation, 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

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 well-normed 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 poses 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 to 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 well-developed, psychometrically sound assessments will greatly enhance the likelihood that accurate and valid information can be obtained.

REFERENCES

American Educational Research Association (AERA), American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: AERA.

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

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

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.

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

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

Chakrabarti, S., & Fombonne, E. (2001). Pervasive developmental disorders in preschool children. *Journal of the American Medical Association, 285,* 3093–3099.

Cohen, J. (1988). *Statistical power analysis in the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

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.

Constantino, J. N., Hudziak, J. J., & Todd, R. D. (2003). Deficits in reciprocal social behavior in male twins: Evidence for a genetically independent domain of psychopathology. *Journal of the American Academy of Child and Adolescent Psychiatry, 42,* 458–467.

Constantino, J. N., & Gruber, C. P. (2005). *Social Responsiveness Scale.* Los Angeles: Western Psychological Services.

Constantino, J. N., Gruber, C. P., Davis, S., Hayes, S., Passanante, N., & Przybeck, R. (2004). The factor structure of autistic traits. *Journal of Child Psychology and Psychiatry, 45,* 719–726.

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.

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.

DiLavore, P., Lord, C., & Rutter, M. (1995). Pre-linguistic autism diagnostic Observation Schedule (PL-ADOS). *Journal of Autism and Developmental Disorders, 25,* 355–379.

Goldstein, S., & Naglieri, J. A. (in press). *Autism Rating Scale.* Toronto: Multi-health Systems.

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

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., Rutter, M., Goode, S., Heemsbergen, J., Jordan, H., Mawhood, L., & Schopler, E. (1989). Autism Diagnostic Observation Schedule: A standardized observation of communicative and social behavior. *Journal of Autism and Developmental Disorders, 19,* 185–212.

Lord, C., Rutter, M., DiLavore, P. C., & Risi, S. (2002). *Autism Diagnostic Observation Schedule*. Los Angeles: Western Psychological Services.

Naglieri, J. A., & Das, J. P. (1997). *Cognitive Assessment System interpretive handbook*. Itasca, IL: Riverside.

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.

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

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

Rutter, M., Le Couteur, A., & Lord, C. (2003b). *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., Reichler, R. J., & Renner, B. R. (1988). *Childhood Autism Rating Scale*. Los Angeles: Western Psychological Services.

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.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children—Fourth Edition*. San Antonio, TX: Psychological Corporation.