

CHAPTER 28

*Diagnostic Instruments in Autistic
Spectrum Disorders*

CATHERINE LORD AND CHRISTINA CORSELLO

The development of diagnostic instruments in the past 30 years is an example of the interplay between clinical and research needs in the field of autism. When judged from field trials of diagnostic criteria (Volkmar et al., 1994), autism is one of the most reliably diagnosed disorders in child psychiatry. However, many diagnostic aspects of the disorder provide unique challenges, as well as raising issues shared with other childhood onset disorders. In this chapter, first general and then specific issues pertaining to designing and selecting instruments for diagnosis and measurement of core features of autistic spectrum disorders (ASDs) are considered. A brief historical review of some of the first standardized instruments used for diagnosis of autism is next, followed by short descriptions of some of the most common instruments used in diagnosis and measurement of the features that define ASD. The chapter concludes with information about the use of instruments for specific purposes, such as measuring change, ending with a general discussion. Because the emphasis of the chapter is on issues pertaining to the design and selection of measures, sections on individual instruments are not intended to be comprehensive. See review articles by Parks (1988)

and Teal and Wiebe (1986), as well as original works cited in the text, for further information.

**GENERAL ISSUES IN DIAGNOSIS OF
AUTISTIC SPECTRUM DISORDERS**

Autism and other pervasive developmental disorders (ASDs) are associated with a broad range of intellectual and language skills, particularly across time. This range affects the way that the disorder's defining symptoms are manifested. Because ASDs typically begin when children are infants or toddlers and continue into adulthood, precise identification of well-defined behaviors that are necessary and sufficient to diagnoses across developmental levels is a complex task (Lord, Pickles, DiLavore, & Shulman 1996; Volkmar et al., 1994). For example, although deficits in simple pretense and elicited imitation are typical of most children with autism at certain points in development, these deficits do not necessarily discriminate autism from other disorders at either very basic levels of development (i.e., age equivalents of under 12 months; Charman et al., 1998) or at much more sophisticated levels of development (i.e., very high-functioning verbal adults; Happé, 1995).

Appreciation is expressed to NICHD (U19HD35482) through the Collaborative Program for Excellence in Autism (CPEA) and NIMH (R01MH066496) that provided support to the authors during the preparation of this manuscript and to Colleen Hall, Kaite Gotham, Daniel Karstofsky, and Amanda Edgell, who helped in the preparation of this chapter.

The challenge presented by changes in development in autism is similar to issues that affect the measurement of general intellectual development in all children. In the case of general intelligence testing, however, years of investigation, access to large populations, and population samples of normative data have allowed the development of instruments such as the Wechsler tests (WISC-IV; Wechsler, 2003), WAIS-III (Wechsler, 1997), WPPSI-III (Wechsler, 2002). These tests contain different tasks for children and adults at different levels. Standard scores are computed according to small gradations in age. In ASDs, with the exception of the revised Autism Diagnostic Observation Schedule (ADOS: Lord, Rutter, DiLavore, & Risi, 1999; Lord, Risi, et al., 2000), such grading has not yet been attempted, and may not be feasible, given the incidence and variability of the disorders.

In addition, while cognitive tests use chronological age and population demographics to define what is "average," identifying the "average" child with autism is much more complicated, particularly with small samples. Using large samples who have not been systematically assessed or recruited according to epidemiological standards may also lead to unrepresentative scores (see Ozonoff, South, & Miller, 2000). Eventually, pooled research samples that result in very large sample sizes and/or methods such as latent class analyses may be helpful in this endeavor (Mahoney et al., 1998). In the meantime, studies that explicitly compare distributions of different samples (e.g., Szatmari et al., 2002) provide important information about the consistency of diagnosis across populations.

In addition, as discussed in more detail later in this chapter, issues arise about how to best define comparison group to autism in order to generate appropriate norms. Providing normative data based on chronological age, as is done for most well-known general intellectual assessments, is not sufficient, because ASDs are often, but not always, associated with mental handicap. Thus, differences obtained between mentally handicapped children with autism and chronological-age matched nonautistic children who are not mentally handicapped may be attributed to autism, mental handicap, or both. On the other hand, the generation of norms

based on all combinations of chronological age and level of mental handicap is not feasible without very large samples, and sometimes not even then (e.g., identifying infants with mild idiopathic mental retardation may be impossible).

A further factor is language delay. Even when level of mental handicap is addressed through a research design, children with autism-related disorders often (with some notable exceptions) show more severe language delays than other children of equivalent nonverbal level. Any diagnostic instrument that relies heavily on behaviors associated with receptive or expressive language competence must take this into account (Lord, Storoschuk, Rutter, & Pickles, 1993). However, exactly how to do so becomes a complex decision (Happé, 1995; Hobson, 1991). Trying to control for language delay may also "control for" autism itself. It may result in comparisons that are invalid for other reasons (e.g., comparing 2-year-olds with autism to nonhandicapped 8-month-olds of equivalent receptive language skill).

In addition, the relationship between autism and language impairment is complicated by the fact that the expressive language of individuals with no or very little spontaneous speech may not show as many abnormalities as the language of more verbally fluent persons with autism. This relationship affects attempts to quantify severity in any additive way. Thus, in the Autism Behavior Checklist (ABC; Krug, Arick, & Almond, 1980b) and in the Autism Diagnostic Interview-Revised (ADI-R; Rutter, Le Couteur, & Lord, 2003), both described later, an abnormality score is computed by adding the number of ways in which a child or adult's language is unusual (e.g., pronoun reversal, delayed echolalia, neologisms). This strategy results in individuals with more complex language scoring as more abnormal than individuals who cannot speak (Miranda-Linne, Fredrika, & Melin, 1997; Rutter et al., 2003). A recent factor analysis carried out on the ADI-R (Lord, Rutter, & Le Couteur, 1994; Tadevosyan-Leyfer et al., 2003) credited nonverbal children with maximum scores of severity on verbal items. This resulted in nonverbal children scoring as most severe on a hierarchy of language items, overlapping with children with the most sophisticated language and many abnormalities; not a result that is very meaningful

or interpretable. The ADI-R attempts to avoid this problem by having separate domain scores for verbal and nonverbal communication; however, this strategy is not ideal for researchers who need a single overall severity score.

In general, classification systems and diagnostic instruments for ASDs have been most accurate in addressing autism in somewhat verbal, mildly to moderately mentally handicapped school age children. Classification systems and diagnostic instruments decrease in interpretability the farther one moves from this group (Lord & Bailey, 2002; Lord et al., 1996). Unfortunately, diagnostic instruments are most needed for children and adults who do not fall within this most easily recognized prototype. As discussed later, it is important that consumers who use diagnostic instruments take into account the biases that an instrument shows for populations who fall outside the most commonly studied group of children with autism, such as children with nonautism ASDs, such as Asperger's Disorder and Pervasive Developmental Disorders, Not Otherwise Specified (PDD-NOS). The difficulties are less relevant for Rett Syndrome and Fragile X, because these disorders have biological markers, however, questions remain when children with these disorders meet standard diagnostic criteria for autism.

Issues in Selecting the Appropriate Focus and Level of Analysis

An alternative to organizing a diagnostic instrument around very specific behaviors is to develop measures of broadly defined deficits, such as impairments in social reciprocity or circumscribed interests that are relevant to the behaviors of individuals across a range of chronological ages and developmental levels. However, answering questions about these broad conceptualizations may be difficult for naive observers, such as nonexpert clinicians (Volkmar et al., 1994) or parents (Schopler & Reichler, 1972). This seems especially true in diagnoses of young children (see Charman et al., 1998; DiLavore, Lord, & Rutter, 1995; Lord et al., 1993), for whom it may be difficult to disentangle well-coordinated social behaviors produced as part of familiar, physical routines from spontaneous, socially motivated

interactions. For example, in a study comparing parent report in a structured interview to direct observations, good agreement across the two methods for the occurrence of abnormalities emerged for only 3 of 16 items taken from *DSM-III-R*: abnormal social play, stereotyped body movements, and restricted range of interests (Stone & Lemanek, 1990). Differentiation for adults between deficits specific to autism and those associated with any severe, chronic psychiatric disorder that drastically limits social contact and everyday opportunities, also becomes more difficult (Rutter, Mawhood, & Howlin, 1992; Volkmar et al., 1994).

Parent and child reports are not interchangeable. This issue is most relevant to high-functioning older children, adolescents, and adults with autism and ASDs who can be asked to describe their own symptoms and concerns. For certain behaviors, parent report may be more valid and reliable over time (e.g., reports of friendships, development of play; Lord et al., 1989); for others, either direct observation (such as of very young children with autism; Lord, Cook, Leventhal, & Amaral, 2000) or self-reports, such as for mood and interest in the opposite sex (Howlin, Mawhood, & Rutter, 2000; Mawhood, Howlin, & Rutter, 2000), may be more accurate indicators. In other areas of developmental psychopathology, with a few notable exceptions (e.g., self-reports of anxiety or depressive feelings), informant accounts have often been better discriminators than alternative methods (Bird, Gould, & Staghezza, 1992).

Using multiple sources may address some of these issues by helping to place diagnostic information in developmental and social contexts. For example, if a child appeared fascinated by pencils during an observation, a parent's account of his fascination with stick-like materials at home would be important in evaluating whether this was a consistent focus or a brief interest. Information about a history of very limited social interaction beginning in early childhood can place reports of social isolation into context for an adult client. From the reverse perspective, observation of how a child responds when a parent is asked to call his name may be a helpful complement to a parent's description of the child's response to family members' attempts to get his attention at home. Ideally, diagnostic instruments would

maximize use of direct observations and parents' and teachers' descriptions, while getting broader information directly from individuals with ASD without requiring them to draw inferences that they often do not have the knowledge to make (e.g., about the nature of autism and the applicability of that term to themselves). However, how to best combine information from multiple sources is not obvious (Kraemer, 1992; Offord et al., 1996). For example, one method of quantifying severity might be to consider information from different sources as separate repeated measures of a hypothetical construct, such as qualitative impairments in social interaction (Grinager, Cox, & Yairi, 1997).

Instruments also differ in the degree to which they emphasize the presence of observable abnormalities or the absence of normally developing features. Sometimes this distinction is arbitrary, as in descriptions of the use of gaze by children with autism as either "unusual eye contact" or "failure to use gaze to regulate social interaction in subtle ways." The former describes the presence of an abnormality and the latter describes the absence of a prosocial behavior. In young children with autism, the absence of behaviors such as eye contact, smiling, and social responses, may be more specific and more predictive of outcome than abnormalities (Lord, 1995; Venter, Lord, & Schopler, 1992). It is also more highly correlated with chronological and cognitive age (Tadevosyan-Leyfer et al., 2003). For other diagnostic features, the presence of clear abnormalities and the absence of normal development may be strongly related, but the two perspectives may not necessarily be the same. For example, developmental and behavioral intervention studies would suggest that the presence of unusual preoccupations and restricted interests is associated with the absence of early social play. If a child is taught developmentally appropriate play skills, he will show fewer stereotyped behaviors (Schopler, 1976); however, he may still have restricted interests. To our knowledge, this assumption has not been directly tested outside of evaluations of specific interventions.

Even though the two approaches (computing the presence of abnormalities and determining the number of absences of prosocial

features) are clearly related, they have somewhat different implications for diagnostic instruments. Social-communicative features of autism tend to be described in terms of absences, while oddities in interests and behavior, as well as a few specific characteristics of language (e.g., stereotypic speech) tend to be described in terms of the presence of abnormalities. When they occur, odd behaviors, such as hand and finger mannerisms or repeated smelling of objects, may be more striking and obviously abnormal than the lack of typical development in a particular area. However, such obviously abnormal behaviors, even if a child or adult engages in them frequently at home or school, may not always occur during a relatively brief observation. For example, in one study, only 60% of verbal, mildly mentally handicapped adolescents with autism and 35% of very high-functioning, verbal adolescents with autism exhibited clearly observable repetitive behaviors during a half-hour structured observation, though all of these individuals were described by their parents as engaging in such behaviors at home on a regular basis (Lord et al., 1989). None of the language and chronological-age matched mentally handicapped and normally developing adolescents exhibited these behaviors during the observation. The *presence* of these behaviors during an observation was important diagnostically, but the *absence* during that one observation was not interpretable. As already noted, there is reason to believe that such abnormalities may be less directly related to clinical outcome than are social impairments and more broadly based aspects of communication (Cox et al., 1999; Venter et al., 1992). Nevertheless, brief descriptions of clearly abnormal behaviors, particularly sensory reactions to environmental stimuli, are more amenable to checklists and screening measures (Krug et al., 1980b; Rimland, 1971) than longer-winded descriptions of subtle differences in nonverbal social behaviors, though the abnormal behaviors may be less indicative of outcome and of diagnoses made by experienced clinicians than other measures.

It is important to remember that, in a diagnosis, the diagnosticians tend to find what they look for or ask about. That is, the content and the nature of the behaviors that are observed

(or described) and the content and the nature of the ways in which they are reduced or “coded” affect the end product of diagnosis. Scales that employ linear approaches to scores (e.g., using a single total) with a single cut-off more easily quantify examples of dysfunction, but also are more likely affected by factors outside autism, most notably co-occurring mental retardation, than are instruments that require thresholds in different areas. Scales that require meeting of multiple thresholds are tied to specific classification systems and the theories that underlie them (e.g., *DSM IV* and *ICD-10*). Thus, they may underestimate cases because of requirements for distribution of scores or because the system is not quite correct (Cox et al., 1999; Hepburn, John, Lord, & Rogers, 2003; Lord, 1995; Pilowsky, Yirmiya, Shulman, & Dover, 1998).

For example, one study showed that both the Childhood Autism Rating Scale (CARS; Schopler, Reichler, & Renner, 1988) and the Autism Diagnostic Interview-Revised (Lord et al., 1994) were concordant with clinician’s judgments of diagnosing autism in children at age 3 (Lord, 1995). Both were less accurate for children 2 years or younger, but for somewhat different reasons. The CARS consistently over-diagnosed nonautistic mentally handicapped children as having autism at age 2; CARS diagnoses of these children became more accurate by age 3, but were still less specific than has typically been reported for older children. The ADI-R was more accurate than the CARS with the nonautistic children at 2, but like the CARS it was over-inclusive for mentally handicapped and/or language delayed children. The ADI-R also failed to diagnose autism in about 10% of 2-year-olds who later met formal diagnostic criteria for the disorder because their parents did not report sufficient abnormal repetitive behaviors or abnormalities in language. Agreement between the ADI-R and CARS was in fact quite high; the difference was whether a simple total or thresholds across several domains (i.e., social reciprocity, communication, restricted, repetitive behaviors) were required for a diagnosis.

Similar results were found in another study comparing the ADI-R and CARS with older children (Pilowsky et al., 1998). The ADI-R resulted in good specificity, but poor sensitivity at detecting childhood autism at 20 months

of age (Cox et al., 1999). Furthermore, the ADI-R was not sensitive to other Pervasive Developmental Disorders, such as Asperger’s Disorder and PDD-NOS, when used with 20-month-old toddlers (Cox et al., 1999). Stone found that clinical diagnoses at age two identified children with later stable diagnoses of autism but not of PDD-NOS (Stone, Ousley, Yoder, Hogan, & Hepburn, 1997; Stone et al., 1999). As we discuss later, decisions of which approach is most appropriate may differ depending on the needs of the clinician or researcher and the developmental level of the child or adult who is assessed.

Implications of Information from Other Areas of Research for Diagnostic Instruments

Without a well-established biological marker, decisions about classification of autism and ASDs have often been based on the need to identify appropriate populations for services and research, rather than empirical bases (American Psychiatric Association, 1994; Volkmar et al., 1994; Wing & Gould, 1979). Though eventually, neurobiological factors may result in a re-sorting of diagnoses in autism/PDD, biological heterogeneity is expected within and among the spectrum disorders. Thus, we will be dependent on descriptions of social and other behaviors for some time. Yet, the behavioral boundaries between autism and other disorders in the spectrum, such as PDD-NOS and Asperger’s Disorder, are not clearly defined, particularly when changes with development are taken into account (see Ghaziuddin, Tsai, & Ghaziuddin, 1992). Information, such as developmental trajectories and clustering of symptoms, that arises out of studies of diagnostic instruments may influence classification systems in the near future (Lord et al., 1996; Mahoney et al., 1998; Szatmari, Archer, Fisman, Streiner, & Wilson, 1995). The expectation is that diagnostic instruments may and should continue to change as more information is acquired.

Furthermore, priorities for the results of diagnoses may be different for clinical and research purposes. Clinical diagnoses offer families access to general information about their children. A clinical diagnosis is often the entry point to services. Service providers may

use a diagnosis to allocate limited resources, whereas a priority for families and diagnosticians is to ensure that children or adults are not being excluded from appropriate services because of a particular label or classification (Wing & Attwood, 1987).

Researchers often prefer narrow diagnoses. Narrower formulations provide better cross-site reliability, eliminate outliers, and reduce overlap with control groups. Narrower diagnostic categories may reduce the likelihood of false positives. On the other hand, researchers seek populations of particular sizes and are interested in maximizing the number of participants who meet their criteria. All of these forces affect the goals addressed by diagnostic instruments and the ways in which they are used.

There is an urgent need for instruments to address diagnoses beyond autism, particularly ASDs, such as PDD-NOS and Asperger's Disorder. In part, the absence of replicable, reliable, and valid instruments in this area is related to the absence of clear diagnostic criteria for these disorders (Sponheim, 1996; Szatmari et al., 2002). A lack of empirical data affects the ability to discriminate these disorders both from autism and from disorders outside the autism spectrum (e.g., severe attention deficit; severe communication impairment), which in turn affects the development and the operationalization of these criteria.

There are numerous sets of diagnostic criteria for ASDs, especially for Asperger's Disorder, that suggest conceptualizations for them (Volkmar & Klin, 2001; Szatmari, 2000; Tantam, 2000), but that do not directly address the overlap with autism. In contrast, *DSM-IV* and *ICD-10* criteria define Asperger's Disorder purely in terms of its relationship with autism, but provide little conceptualization (*DSM-IV-TR*). Moreover, conceptualizations exist for disorders such as schizoid disorder and nonverbal learning disabilities; but without clear indications of their relationship with autism. As they are for Asperger's Disorder, *DSM-IV* criteria for PDD-NOS and *ICD-10* criteria for atypical autism are based solely on the basis of just missing autism criteria.

Developing standardized assessment instruments for Asperger's Disorder is particularly tricky because there is little consensus in how to define the disorder, little consistency in

the manner in which current criteria in the *DSM-IV* and the *ICD-10* should be applied, and little agreement as to whether the diagnosis is distinct from autism or is a subtype of autism (Klin, Pauls, Schultz, & Volkmar, in press). Several different definitions for Asperger's Disorder are currently used to make the diagnosis, including Gillberg's criteria, Szatmari's criteria, Tantum's criteria, and the criteria listed in *DSM-IV* and *ICD-10* (Leekam, Libby, Wing, Gould, & Gillberg, 2000). Some authors have reported that it is difficult, if not impossible, to diagnose Asperger's Disorder given the current diagnostic criteria, which requires that autism is excluded prior to making a diagnosis of Asperger's Disorder (Miller & Ozonoff, 1997; Szatmari et al., 1995).

There are also different opinions of what criteria to use to determine if Asperger's Disorder should be considered distinct from autism or not. Klin and his colleagues broach the need for a greater body of research on the validation of the syndrome (Klin, Sparrow, Marans, Carter, & Volkmar, 2000). Szatmari et al. (1995), on the other hand, argues that the decision should be based on clinical usefulness, taking into account course, response to treatment, and prognosis. Szatmari et al. indicates that the first priority should be to determine if there is a meaningful distinction between Asperger's Disorder and autism and a second priority should be to determine if there is a distinction between Asperger's Disorder and other related but also not well-defined groups (such as nonverbal learning disability). One of the difficulties is that different criteria for Asperger's Disorder change not only the individuals who receive that diagnosis, but also whom is then diagnosed with PDD-NOS and autism (Klin et al., in press).

At this point, there are very few instruments with extensive reliability and validity studies available to aide in the diagnosis of Asperger's Disorder. The few that are available are described in this chapter. Most researchers continue to modify instruments designed for diagnostic purposes for autism in research studies on Asperger's Disorder, particularly as there is still controversy as to whether it is a distinct syndrome.

Two Asperger's Disorder algorithms were developed for the DISCO (Leekam et al.,

2000), one of the algorithms was based on Gillberg's criteria and the other on *ICD-10*. In the study by Leekam et al. (2000), 91 (45%) of the subjects met criteria for Asperger's Disorder using the algorithm based on Gillberg's criteria, while only 3 (1%) met criteria based on the *ICD-10*. In most cases, this was at least partially due to the *ICD-10* criteria requiring normal language development prior to 3 years and age appropriate self-help or adaptive skills, or curiosity. This supports the difficulty in making the diagnosis based on current criteria as it is set forth in the diagnostic manuals. The authors of this article admit that they may have interpreted the *ICD-10* criteria more strictly than was intended. Also of importance was the finding that all 91 of the children who met algorithm cut-offs for Asperger's Disorder based on Gillberg's criteria also met *ICD-10* criteria for autism or atypical autism, again, highlighting the issue of the overlap between Asperger's Disorder and autism. Similar findings have been reported by numerous other researchers, including Ozonoff et al. (2000), Szatmari et al. (1995), and Klin, Volkmar, Sparrow, Cicchetti, and Rourke (1995).

Other instruments have tended to yield the same results. This is truly an unfortunate cycle: without reliable diagnostic criteria and measures, empirical findings are very difficult to interpret (Klin et al., in press; Sponheim, 1996). Without empirical data about the course and characteristics of nonautism pervasive developmental disorders, attempts to differentiate between these disorders and autism will not be effective. Data from genetic and family studies, as well as other neurobiological approaches, may make this task easier, but the results are also affected by instrumentation. Thus, researchers must arrive at working agreements that allow them to proceed in a reliable fashion.

In the face of these difficulties, autism as a field has the strength of its intense research history and the benefit of research teams from around the world investigating similar questions. Descriptive and experimental research have offered solutions to some of these difficulties, such as identifying developmentally meaningful behaviors—joint attention, theory of mind, response to name—that discriminate autism from other disorders at various points in development. It offers the promise of other

knowledge, from new statistical techniques to neuroimaging to molecular genetics. As perspectives on autism have shifted with new theories and empirical findings, strategies and content of instruments used for its diagnosis have also shifted in numerous ways. However, in the newer instruments, roots can almost always be traced to strategies begun in earlier work. Science offers clinicians the opportunity to learn from accumulated knowledge and empirical testing of hypotheses.

Psychometric Issues

The American Psychological Association (APA) has issued guidelines for the development of psychometric instruments in the United States. A number of factors affect the psychometric appropriateness of an instrument. These issues are raised as they apply to the question of diagnostic instruments for the autism spectrum in general, followed by more specific discussions of selected instruments. Selected standards from these guidelines are presented in Table 28.1 (reliability) and Table 28.2 (validity). Many diagnostic instruments in autism/PDD, as noted, have addressed some of these issues, but few or none have addressed all of them. In part, this lack of information is understandable because of difficulties in achieving sufficiently large well-documented samples; in part, it reflects the limited history of instrument development in autism.

Reliability

Reliability, which is the degree to which a score or decision is free from errors of measurement, requires assessment in a number of forms, including across raters, across time, and within an instrument. Often the term reliability is used to describe these separate aspects of the stability of the results of an instrument as if they were interchangeable. However, this is not the case. For example, the degree to which different raters concur when using the same instrument cannot be determined by measuring the internal characteristics of a test. The internal consistency (i.e., the degree to which different items on a scale measure the same concept) of an instrument can be quite high, even though its inter-rater

TABLE 28.1 Reliability and Errors of Measurement: Issues Related to Diagnosis of Autistic Spectrum Disorders

1. For each total score, subscore, or combination of scores that is reported, estimates of relevant reliabilities and standard errors of measurement should be provided in adequate detail to enable the test user to judge whether scores are sufficiently accurate for the intended use of the test.
2. The procedures that are used to obtain samples of individuals, groups or observations for the purpose of estimating reliabilities and standard errors of measurement, as well as the nature of the populations involved, should be described.
3. The conditions under which the reliability estimate was obtained and the situations to which it may be applicable should be explained clearly.
4. Coefficients based on internal analysis should not be interpreted as substitutes for alternate-form reliability or estimates of stability over time unless other evidence supports that interpretation in a particular context.
5. Where judgmental processes enter into the scoring of a test, evidence on the degree of agreement between independent scorings should be provided.
6. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported for score levels at or near the cut score. For dichotomous decisions, estimates should be provided of the percentage of test takers who are classified in the same way on two occasions or on alternate forms of the test.

Selected and adapted from *Standards for Educational and Psychological Testing*, by AERA, APA, NCME, 1985, Washington, DC: American Psychological Association.

reliability is low. In a disorder such as autism that is defined by a pattern of difficulties across several areas (i.e., communication, social interaction, behavior), internal consistency in a scale is a worthwhile endeavor, but does not have the same meaning as in a scale that is not designed to describe a pattern of related, but different, deficits.

In the past, reliability estimates were often reported as correlations. A correlation measures whether the rankings of different individuals are similar across different raters. The

difficulty with correlations is that the absolute scores of raters can be quite different, resulting in different diagnoses, even though they are highly correlated. That is, if one rater rated all participants relatively high and another rater rated the same participants relatively low and the raters had the same rankings of participants, the correlation of the two raters' scores would be high. If diagnosis is based on exceeding a certain threshold, the fact that the rankings of the raters agreed would not prevent the scores from resulting in different diagnoses for

TABLE 28.2 Validity: Issues Related to Diagnosis of Autistic Spectrum Disorders

1. Evidence of validity should be presented for the major types of inferences for which the use of a test is recommended.
2. If validity for some common interpretation has not been investigated, that fact should be made clear, and potential users should be cautioned about making such interpretations.
3. The composition of the validation sample should be described in as much detail as is practicable.
4. When criteria are composed of rater judgments, the relevant training, experience, and qualifications of the experts should be described.
5. When a test is proposed as a measure of a construct, that construct should be distinguished from other constructs. Evidence should be presented to show that a test does not depend heavily on extraneous constructs. If evidence indicates that a criterion measure is affected to a substantial degree by irrelevant factors, this evidence should be reported.
6. When criteria are composed of rater judgments, the degree of knowledge that raters have concerning rater performance should be reported. The training and experience of the raters should be described.
7. If specific cut scores are recommended for decision making (for example, in differential diagnosis), the user's guide should caution that the rates of misclassification will vary depending on the percentage of individuals tested who actually belong in each category.

Selected and adapted from *Standards for Educational and Psychological Testing*, by AERA, APA, NCME, 1985, Washington, DC: American Psychological Association.

the same client. Thus, while correlations provide an important index of the relationship among scores, they are not sufficient to show agreement when cut-off scores are used to make categorical judgments about diagnoses.

In place of correlations, many investigators now employ measures of percent of agreement between pairs or larger groupings of raters. An agreement must be defined at a level commensurate with the aims of the instrument. It may be exact agreement or agreement within a certain number of points, depending how scores are to be used. Clinicians and researchers can then evaluate the frequency with which their coding agrees with that of another person for a given individual. There are no set standards for levels of agreement, but generally, in self-report and interview studies, researchers have been able to achieve 90% or greater agreement on individual categorical measures and at least 80% on individual observational codes, with greater agreement for pooled or summary scores.

Item-level inter-rater agreement is very important when an instrument is being developed because it allows for experimentation with which items yield the most valid scores. Many of the most well-known assessment instruments (i.e., the Wechsler tests, the Vineland Adaptive Behavior Scales) do not have this level of inter-rater reliability because they rely on total or domain scores and because the internal consistency of these domains or the total are well documented. In the field of ASD, because of gradually changing conceptualizations, recent instruments have actually aimed for the establishment of more specific reliability among raters in order to retain the flexibility to rework scoring systems as different diagnostic frameworks emerge.

The difficulty with using percent agreement as a metric is the role of chance. If there is a high frequency of extreme scores without much variation within different populations (e.g., almost all zeroes for nonautism or high scores for autism), correlations and percent agreement among raters can be quite high because of the likelihood of agreement based on using the extreme scores, without attention to individual differences. That is, having seen a child's performance on the first item of the test, a rater might predict that, because the child looked quite autistic on the first item, he will receive

high scores on all further measures of abnormality. Having seen a typically developing child's behavior on the same first item, a rater might predict, based on the child's "normal" reaction to the first task, that she will receive "normal" scores on other items. If there is little variation across tasks and little overlap across populations, two raters might get better agreement using this strategy than by actually observing and coding the behaviors of the individual children. Specific statistics, called kappas (Cicchetti & Sparrow, 1981), allow some control of this phenomenon. However, no simple answer addresses all of these problems. Although kappas control for chance, they are sensitive to distributions and so, as with any statistic, must be interpreted in light of other information. Another strategy using reliability coefficients does not address the intersection between individual participants and individual raters, but allows quantification of the effects of each separately (Mundy, Sigman, Ungerer, & Sherman, 1986). This statistic tests whether scores are more affected by individual differences in children than by differences among raters. However, if there are large individual differences among children, finding that these differences exceed those among raters may not guarantee strong reliability.

These issues illustrate the importance of the nature of the samples on which psychometric analyses are conducted. Autism affects individuals across the lifespan who have a range of language and cognitive skills. If samples are not well matched and not relevant to the clinical or research contexts in which the instrument will be used, there will generally be little overlap in scores (e.g., if children with autism are compared to typical children). If instruments are developed only using very easily discriminable populations, documentation of reliable ratings will be difficult to achieve when statistics that take distributions into account are employed, although they may look good in terms of absolute agreement. When reliability estimates are presented only for totals, even when subscales are described and intended to be used, clinicians or researchers who want to base interpretations on specific items or subscales cannot do so. It is important that test users interpret their results within the context of the information that is available.

Sometimes the reverse is the case. Researchers may present detailed psychometric data for items, but not present reliability for the diagnostic categorization for which the scale is intended. This is particularly problematic for ASDs. It is not difficult to find an instrument that identifies more abnormal behaviors occurring in ASD than in typical development. However, seldom is this the goal of an instrument. To be useful diagnostically, instruments must discriminate children with autism or ASD from nonautistic severely mentally handicapped, language-impaired children. Because it is often difficult to set a threshold that includes children with mild autism identified as such and excludes nonautistic severely mentally handicapped children, consistency across raters and across time with which an individual falls in or out of the category of autism or ASD must be measured directly.

The issue of test-retest reliability in autism is complex. Changes in behavior due to development would be expected if administrations were separated by substantial amounts of time. Some learning may occur within the testing situation that affects a child's behavior if he or she is asked to carry out the same actions again. This is different than error in measurement, but still must be taken into account. In some cases, previous administration of an instrument (i.e., practice) may affect its scores or interpretation. For example, in the Autism Diagnostic Observation Schedule (ADOS, Module 1; Lord, Risi, et al., 2000), young children are taught a routine of bringing a balloon to the examiner if they do not do so spontaneously. If they are presented with the same task several weeks later, they may respond differently because of learning, not because of error in measurement. However, the examiner still needs to code the behavior he or she sees. Ideally, information about stability and expected changes across multiple administrations should be available for all instruments.

For diagnostic instruments, this information must be presented at the level of each individual's score and resulting diagnosis. Just because a task or instrument has been used in many studies, it cannot be assumed that it is reliable on an individual level at a standard appropriate for diagnostic work. Many experimental studies in psychology and psychiatry are primarily con-

cerned with identifying group differences and so do not address issues at an individual level in much detail, if at all. For example, two studies reported substantial intra-individual variability across tasks and time in standard tasks used to assess theory of mind (see Chapters 41 and 42 in this book for a discussion of this concept) in autism (Holroyd & Baron-Cohen, 1993; Mayes & Zigler, 1992). While group effects on false belief tasks have had a major impact on the conceptualization of social-cognitive deficits in autism, and have been replicated across studies internationally, in neither of the recent studies were the results of the tasks sufficiently replicable within individuals to meet reasonable clinical standards for classification.

An important aspect of reliability is specification of exactly how and under what circumstances diagnostic instruments are to be used and how they are to be scored. Sometimes procedures reported in journal articles are described so briefly that it is difficult to determine what exactly was done and who did it. Differences in procedures, such as whether or not coding is carried out live or from videotape, whether interviews are done face to face or on the telephone, or how experienced in autism the raters are, may result in differences in scoring (Sanchez et al., 1995; Volkmar et al., 1994). It is helpful for users of instruments to know how, as "consumers," they might improve and evaluate their own reliability with an instrument.

In studies of reliability and validity, raters should be unaware of children's diagnostic categories or of scores on other diagnostic instruments, unless this information would typically be available prior to use of the instrument. If other information is assumed to be a critical part of the use of the instrument, this needs to be stated clearly as part of the procedures. For example, for the ADOS, general information about a participant's likely level of expressive language is crucial in selecting the appropriate module and so is considered part of the assessment. How this information is used is specified in the manual. In addition, description of the training required for a rater and the circumstances of the training and the administration are critical aspects of reliability.

Another factor to be considered in autism and ASDs is parents' awareness of their

child's diagnosis. That is, in many research samples, parents of previously diagnosed autistic children are well versed in the characteristics of autism and how their children fit into the diagnostic scheme. Several recent studies have shown excellent agreement between questionnaires (i.e., the Social Communication Questionnaire [SCQ]; Rutter, Le Couteur, & Lord, 2003) and interview formats of similar items (i.e., the SCQ and ADI-R; Bishop & Norbury, 2002; Chakrabarti & Fombonne, 2001; Le Couteur, Lord, & Rutter, 2003; Lord et al., 1994; Vrancic et al., 2002 [Spanish SCQ by telephone]). However, if a parent report instrument is intended to be used in initial diagnoses, then it is appropriate that it is shown to be reliable and valid with caregivers who have not yet received formal diagnoses.

Validity

Validity is the most important aspect of a diagnostic instrument. Validation refers to the degree to which other evidence supports inferences drawn from the scores yielded by the diagnostic instrument. Thus, how validity is best measured is inherently related to the uses for which the instrument is intended.

Validity is often grouped into categories of content, construct, and criterion-related evidence. For the diagnosis of ASDs, questions of construct validity are related to those that underlie the diagnostic framework on which the instrument is based. For example, the ADI-R uses a concept of social reciprocity derived from theories of autism (see Lord & Bailey, 2002). It is operationalized in terms of specific questions to parents and caregivers about behaviors such as joint attention, shared enjoyment, comforting, and friendship. Data from studies of the ADI and ADI-R (Le Couteur et al., 1989; Lord et al., 1996) contributed to the understanding of this construct during preparation of *DSM-IV* and *ICD-10* criteria, along with results of observational studies and field trials, in showing that traditional measures of attachment were not strongly related to other measures of social reciprocity (Lord et al., 1993; Sigman & Ungerer, 1984; Volkmar et al., 1994). A further study suggested that

parental reports on the ADI, of autistic children's responses to separation and reunion (which were intended to be linked theoretically to conceptualizations of attachment) were more highly correlated with their children's communicative competence than the same children's observed responses to separation and reunion in a standardized setting (e.g., during administration of the Pre-Linguistic Autism Observation Schedule or PL-ADOS; DiLavore et al., 1995; Spencer, 1993).

Internal consistency for items within a diagnostic instrument can be used to support the assertion that a test measures a single construct. In ASD, this has meant support for the differentiation of ASDs from other developmental disorders or support for the three domains (social reciprocity; communication; restricted, repetitive behaviors) that define the syndrome. Measures of internal consistency for the most commonly used instruments in the diagnosis of autism (e.g., the ADI-R, the Autism Behavior Checklist or ABC, the ADOS, the Childhood Autism Rating Scale or CARS) have generally been high.

Content validity has to do with the degree to which a sample of items, tasks, or questions in an instrument are representative of a defined domain. In most cases, this domain is autism, either narrowly or broadly defined (i.e., ASD, PDD). For the purposes of this review, content validity is most often defined as the degree to which different instruments represent the diagnostic criteria for ASDs. Many of the instruments reviewed here predated the release of *DSM-IV* and *ICD-10* criteria for autism and so do not correspond to the three-domain approach specified in these diagnostic systems. The exceptions are the ADI-R and ADOS. These are special cases because interpretation of results from the original versions of these instruments, the ADI and ADOS/PL-ADOS, influenced strategies tested in the field trials and the *ICD-10* revisions.

Concurrent aspects of criterion-related validity of instruments have been most commonly addressed in the broad area of ASD by investigating the convergence between diagnostic categorizations yielded by another diagnostic instrument or with clinical judgment. As shown

in Table 28.3, convergent validity for three of the most common diagnostic instruments (ADI, ADOS, CARS) available in English has been quite good. Convergence between the CARS and several other instruments (e.g., the Autism Behavior Checklist; Krug et al., 1993; the Real-Life Rating Scale or RLRs; Freeman, Ritvo, Yokota, & Ritvo, 1986) has been good. Also, as depicted in Table 28.3, all of the diagnostic instruments have been shown to be adequate in identifying clinically diagnosed children with autism, with relatively rare false negatives within a "prototypical" group of mildly to moderately mentally handicapped school age children with autism. There is more variability when instruments are used with younger (Lord, 1995; Lord et al., 1993) and older (CARS; Garfin, McCallon, & Cox, 1988; Piven, Harper, Palmers, & Arndt, 1996) populations, and with higher (Yirmiya, Sigman, & Freeman, 1994) and lower functioning groups (Fombonne, 1992; Lord et al., 1993). This pattern is not unique to the instruments, but reflects general difficulty in application of standard diagnostic criteria to various developmental levels. More detailed information about this issue is discussed next with descriptions of particular instruments.

An even more serious, though less widespread, issue is that of false positives. Instruments differ considerably in the number of studies that include comparison groups. They also differ in the degree to which the comparison groups represent typical populations for whom a diagnosis of autism or ASD might be considered and rejected. Often studies include a comparison group of nonautistic mentally handicapped or language impaired subjects, without sufficient information to determine the degree to which these subjects were comparable in ways other than the characteristics of autism to the autistic individuals. Autism is associated with particularly severe communication difficulties; and it is well established that the triad of deficits that define autism increases in frequency as level of mental retardation increases (Wing & Gould, 1979). Consequently, there is reason to be concerned that, without deliberate stratification, most comparison groups of nonautistic individuals will have markedly lower communication skills, adaptive

abilities, and perhaps even general intellectual skills than autistic participants. Thus, comparisons of such samples, even though they may be representative of the population at hand, could yield differences interpreted as specific to autism that may be more accurately linked to severity of mental handicap or communication impairment (Lord et al., 1993). This is another reason why data concerning the size, the characteristics, and the ascertainment of samples are especially important in evaluating instruments. In addition, more sophisticated statistical techniques, such as latent class analyses and logistic regression, may allow researchers to take into account both positive and negative predictive values within a single metric (though still dependent on adequate samples on which to make comparisons).

Little information concerning predictive validity of diagnostic instruments in autism exists except for a few studies using the ADI-R. Our own follow-up study of 2-year-olds who were referred to a pediatric clinic for an evaluation of possible autism, showed that both the ADI-R and the CARS tended to over-diagnose autism in mentally handicapped children at age 2. This was much less the case by age 3, and was less true for the ADI-R (in part, because of the requirement for a "triad" of deficits) than the CARS. On the other hand, Cox et al. (1999) found ADI-R diagnoses, when the threshold in repetitive behavior was not required, to be quite stable from 18 months to 3 years, for a select, higher-functioning group of children identified as having autism with a screening instrument called the CHAT (Baron-Cohen et al., 1992).

A follow-up study from early school age showed that retrospective ADI scores describing behavior at 4 to 5 years of age significantly predicted academic achievement and adaptive scores in adolescence and young adulthood in a group of mildly mentally handicapped to non-retarded autistic individuals (Venter et al., 1992). Social and communication deviance at age 5 made independent contributions, in addition to various measures of expressive and receptive language and nonverbal IQ, to current adaptive skill; whereas the severity of restricted and repetitive behaviors added to the predictive value of verbal and nonverbal predictors of academic achievement.

TABLE 28.3 Currently Available Diagnostic Instruments in Autism

Instrument	Reliability				Validity			General Information		
	Interrater	Test-Retest	Internal Consistency	Construct/Content	Convergent	Discriminant Matched Sample	Published Guidelines for Diagnostic Decision	Subscales	Most Appropriate For	Level of Expertise
Rimland's E-2 form (E-2)	Unpublished	Unpublished	Unpublished	Kanner (1943)	—	Poor	—	—	Screening	Parent check-list
Behavior Rating Instrument for Autistic and Atypical Children (BRIAAC)	S: good	—	S: variable	—	—	Limited	Yes	8	Current observation	Requires training
Real-Life Rating Scale (RLRS)	T: moderate I: marginal	—	T: good S: poor	ASA	CARS ABC	AUT/MR/TYP	—	5	Screening	Minimal
Social Responsiveness Scale (SRS)	T: high	T: high	T: high	DSM-IV	—	TYP/PSY/AUT/ PDD/AS	No	3	Symptoms Severity	—
Pervasive Developmental Disorders Rating Scale (PDDRS)	T: high S: high	T: high S: high	T: good-high S: high	DSM-III-R	ABC	AUT/AD/LDD/ MR/PDD-NOS/ William's Syndrome	No	3	Preliminary stages	Minimal
Children's Social Behavior Questionnaire (CSBQ)	S: variable	T: high S: variable	S: variable	DSM-IV	CBCL ABC	PDD/ADHD/ TYP/PSY/ AUT	No	5	Current symptoms	Minimal
Childhood Autism Rating Scale (CARS)	T: high	—	T: high	DSM-III-R	ABC RLRS ADI	AUT/MR	Yes	4	Targeted screening ^a	Moderate/ video available
Autism Behavior Checklist (ABC)	T: variable	—	T: good S: poor	—	CARS RLRS	—	Yes	2/5	Measuring mal-adaptive behavior	Minimal
Behavioral Summarized Evaluation-Revised (BSE-R)	T: high I: good	—	T: Adequate	—	Rimland E2	AUT/MR/MP	—	—	Symptoms for research	Requires training
Gilliam Autism Rating Scale (GARS)	T: high S: high	T: high S: high	T: high S: high	ASA; DSM-IV	ABC ADPR	AUT/MR/ED/LD (not matched)	Yes	4	Needs further evaluation	Parent check-list

Instrument	S: high I: high	T: good	T: unpublished S: high	DSM-IV ICD-10	CARS ADOS	AUT/MR	Yes	3	Diagnostic clinics/ research across developmental level	Experience, video, or re- quires training
Autism Diagnostic Interview-Revised (ADI-R)	I: variable	—	—	ICD-10; Wing & Gould (1979); Gillberg, Gillberg, et al. (2001); DSM-IV, 1994	—	AUT/LD/LANG DIS	No	4	Educational planning	Requires training
Autism Diagnostic Observation Schedule (ADOS)	I: good T: high	T: Adequate	S: high	DSM-IV ICD-10	ADI-R CCC	MR/LANG DIS/PSY	Yes	3	Research and clinical diagnosis	Experience, video, or requires training
Psychoeducational Profile-Revised (PEP-R)	T: good	—	S: high	DSM-III-R	CARS	TYP	No	4	Intervention recommendations	Experience, video available
Adolescent and Adult Psychoeducational Profile (AAPEP)	S: variable T: high	—	—	—	—	—	No	6	Intervention recommendations	Experience, video available
Communication and Symbolic Behavior Scales (CSBS Behavior Sample)	T: high	T: high S: high	—	—	—	TYP/MR/ASD	No	—	Screening	Minimal, video available
Children's Communication Checklist (CCC)	T: high	—	T: high	—	—	AUT/AD/PDD-NOS/ADHD/LD/LANG DIS	—	5	Identifying pragmatic difficulties	Checklist
Asperger Syndrome (and high-functioning autism) Diagnostic Interview (ASDI)	I: high	I: high	—	Gillberg, Gillberg, et al. (2001); Szatmari (1995); ICD-10; DSM-IV	—	PSY/TYP	No	6	Still in preliminary stages	—
Australian Scale for Asperger's Syndrome (ASAS)	—	—	—	Behavior AS researchers define as AS	—	TYP/ASD/PSY	Yes	5	Screening	Questionnaire

^a Most appropriate for school age children with mental retardation.

Note: All instruments are discussed in detail in text. AD = Asperger Disorder; ADHD = Attention Deficit/Hyperactivity Disorder; AUT = Autistic; ED = Emotionally disturbed; I = Item; LANG DEL = Language delayed; LANG DIS = Language disorder; LD = Learning disabled; MP = Multiple handicap; MR = Mentally retarded; PDD-NOS = Pervasive Developmental Disorder-Not Otherwise Specified; PSY = Psychiatric disorder; S = Subscale; T = Total; TYP = Typical.

DIAGNOSTIC INSTRUMENTS FOR AUTISM

Next, instruments used in the diagnosis of autism and ASDs are discussed briefly following approximate chronological order according to when they were first introduced to the public and according to general categories of method. Descriptions are not meant to be comprehensive; some instruments will be described primarily as examples of kinds of measures or novel approaches. For more detailed information, the reader is referred to specific publications about each instrument or to a chapter by Parks (1988), for many of the older instruments. When several versions of the same or a similar scale have been disseminated, the focus is on the most recent version.

The First Empirically Developed Rating Scales and Questionnaires

The Rimland Diagnostic Form for Behavior-Disturbed Children (Form E-1) was the first widely used scale for the identification of autism (Rimland, 1968). It made an important contribution as a systematic diagnostic assessment that focused on a carefully selected range of symptoms rather than more abstract and inconsistently defined concepts, especially of emotional withdrawal. A revised form, Form E-2 is now scored without charge for parents by the Autism Research Institute in San Diego. Total scores are additive across all questions. The scale is based on the core symptoms defined by Kanner in 1943 and Kanner's belief (Kanner, 1962, as cited in Rimland, 1971) that only a relatively small percentage of children labeled as autistic have "pure" autism.

Many parents have found information from the Autism Research Institute to be helpful. Comparisons with other scales suggest that the diagnosis yielded by the E-2 form is different from those offered by most other instruments. In the original validation study of the Childhood Autism Rating Scale (CARS; Schopler, Reichler, DeVellis, & Daly, 1980; also see below), over 200 children who met autism criteria and another 200 children who did not were all rated on the E-2 form. Only 8 were considered autistic by Rimland using the E-2 form and of those 8, 3 were considered nonautistic on the

CARS. In another study, diagnostic overlap with the Behavior Rating Instrument for Autistic and Atypical Children (BRIAAC; Rutterberg, Dratman, Fraknoi, & Wenar, 1966) was poor (Cohen et al., 1978).

Basic psychometric data and scoring information for the E-2 have not been published in scientific journals (Masters & Miller, 1970). Several studies suggested differences between parent and staff reports using the scale (Davids, 1975; Prior & Bence, 1975) and limited differentiation between children with autism and children with other disorders. While current diagnostic frameworks such as *DSM-IV* and *ICD-10* continue to build on Kanner's original descriptions of autism (Kanner, 1943), the ways in which symptoms are operationalized and weighted have changed substantially. Thus, the E-2 form may serve as most useful to parents who are beginning to familiarize themselves with behaviors associated with autism, rather than as a measure of standard diagnoses of autism or related disorders.

The BRIAAC, is another scale that was created about the same time as Rimland's first diagnostic checklist (Rutterberg, Kalish, Wenar, & Wolf, 1977; Rutterberg et al., 1966). It consists of eight subscales that measure behavior in different areas, yielding a diagnosis of autism. A trained rater completes the scale after substantial observations. The BRIAAC was important historically because it used direct observations of behaviors, defined on the basis of descriptions in case notes (Parks, 1988). Psychometrics were computed on various samples, including at least one study of autistic, mentally handicapped, and normally developing children. Reliability estimates in the form of correlations have consistently been high, though the scoring criteria are complex. More sophisticated estimates of inter-rater or test-retest reliability are not yet published. Results from validity studies have not indicated that diagnostic classifications based on the BRIAAC correspond to those yielded by other instruments or clinical judgment (Cohen et al., 1978). Because it is based only on current observations, the BRIAAC has the potential to be used as a measure of therapeutic effectiveness (Wenar & Rutterberg, 1976), if more up-to-date, rigorous standards for reliability can be met.

Another scale that has been influential in the field of ASDs has been the Handicaps, Behavior, and Skills schedule (HBS) (Wing & Gould, 1978). It was the first widely distributed semi-structured interview for parents and caregivers of children who were mentally retarded or autistic (referred to as "psychotic" at the time). It was used in the Camberwell epidemiological study and, as the source of data for that project, had a significant effect on the understanding of the "triad of impairments" seen in autism and related disorders (Wing & Gould, 1979). The HBS was not a diagnostic instrument, but a "framework for eliciting, systematically, clinical information to be used in conjunction with appropriate psychological tests for assessment and diagnosis" (Wing & Gould, 1978, p. 81). It provided standard questions and topics so that an interviewer could elicit enough information from a parent or caregiver to make an appropriate rating for each item. Formal scoring was mapped onto the Vineland Social Maturity Scale (Doll, 1965). The HBS took several hours to administer and consisted of 31 sections that included questions about both diagnostic and developmental issues. Psychometrics were based on 171 children between 2 and 15 years of age who comprised an epidemiological sample of children with IQs below 50 and/or who were receiving special services who lived in the London borough of Camberwell.

Reliability, judged on the basis of comparisons between pairs of ratings by parents, professional workers, and the authors, averaged from 77% to 81%. Summary ratings across informants and observations in the form of 3-point scales for each section showed near perfect agreement. Indices of association were stronger for the absence of skills than the presence, except for social development. Developmental variables were generally more reliable than ratings of behavioral abnormalities.

One unusual aspect of the reported research was comparisons among professional reports, parent reports, and the authors' direct observations of relevant behaviors. Parents tended to describe their children as more socially and emotionally responsive than did professionals, but to report more stereotyped movements and abnormal responses to sensory stimuli. The more severe the child's impairment, the better was the agreement. The mechanism for combining scores from differ-

ent environments was unique and had the potential for usefulness in documenting changes in behavior. The HBS has now been substantially revised. This revision is discussed later as the Diagnostic Interview for Social and Communication Disorders (DISCO: Wing, Leekam, Libby, Gould, & Larcombe, 2002).

A final scale that was important in the first group of diagnostic instruments emerging in the 1970s was the Behavior Observation Scale (BOS; Freeman, Ritvo, Guthrie, Schroth, & Ball, 1978). It includes ratings of 24 behaviors, carried out in 10-second intervals of a videotaped free-play session. The BOS was the first scale that emphasized the importance of controlling the environment in which a child was observed, as well as standardizing what was observed. It used frequencies of behaviors to differentiate among diagnostic groups. The authors noted that this approach was not completely successful for several reasons. Frequencies of many behaviors were associated with developmental levels as much as diagnosis. In some cases, behaviors that occurred only rarely were very important, suggesting that frequency was a less critical variable than the quality of behavior.

The same authors then developed the Ritvo-Freeman Real Life Rating Scale (RLRS; Freeman et al., 1986) to assess behaviors that characterize autism more accurately, with an emphasis on unusual sensory behaviors. This scale can be used after observation of a 30-minute free-play period. Marginal to adequate reliability was found for individual items with adequate subscale and total inter-rater reliability using kappas (Freeman et al., 1986; Sevin, Matson, Coe, Fee, & Sevin, 1991) for relatively brief samples of behavior coded by raters with minimal training. For a sample of 24 children and adolescents with autism, 7 of 38 items did not occur at all and 4 others were very rare. Inter-rater reliability for another 9 items was not significant (Sevin et al., 1991). On the other hand, the correlation with the CARS total score was .77 for an autistic sample. Three of the five subscales (social relationships, sensory, and language) and the total had adequate to high internal consistency (Sturme, Matson, & Sevin, 1992). No specific cut-offs for diagnosis are provided. Thus, the instrument is primarily useful as a general

index of diagnostic features, and potentially a measure of change, rather than as an independent source of classification.

SCALES THAT MEASURE CORE DEFICITS IN AUTISM SPECTRUM DISORDERS

Social Responsiveness Scale

The Social Responsiveness Scale (SRS; Constantino, 2002), formerly the Social Responsivity Scale, is a questionnaire designed to be completed by an adult, such as a parent or teacher, who observes a child in social situations for the purpose of measuring difficulties in reciprocal social interactions on a continuum (Constantino, Przybeck, Friesen, & Todd, 2000). The questionnaire takes only 15 to 20 minutes to complete and consists of 65 items covering dimensions of communication (6 items), social interactions (35 items), and repetitive and stereotyped behaviors and interests (20 items) associated with ASDs. Each item rates the frequency, not the intensity of a behavior, on a scale from zero (not true) to three (almost always true). The item scores are totaled and result in a severity score along a continuum of difficulties in reciprocity in social interactions (Constantino & Todd, 2000).

Internal consistency of the measure was computed based on teacher completed questionnaires for 195 school children between the ages of 4 and 7 years, resulting in a Cronbach's alpha of .97. All 65 items were retained because reducing the number of items resulted in a reduced ability to distinguish subjects with PDD-NOS from clinical controls. In addition, factor loadings differed between groups of older and younger children. Test-retest reliability has been good with correlations reported between .83 and .88 (Constantino et al., 2004). Inter-rater reliability between parents and teachers ranged between correlations of .73 and .75 (Constantino & Todd, 2000; Constantino et al., 2004) and correlations between parents were also strong ($r = .91$). SRS scores were not related to IQ (Constantino et al., 2004) in one paper, but were in an earlier paper (Constantino, Przybeck, et al., 2000).

Scores on the SRS were significantly higher for children with diagnoses of autism, Asperger's Disorder, and PDD-NOS than for

children in the epidemiological school sample or clinical sample, which was comprised of child psychiatry patients with and without Pervasive Developmental Disorders (PDD). The scores of children with diagnoses of PDDs were approximately 2 standard deviations above the mean of the children with non-PDD psychiatric diagnoses. Approximately 8% of the sample of school children had scores that exceeded the mean of the children with ASDs. While children with PDD-NOS had significantly higher scores than nonautistic children in the clinical sample, overlap occurred between the lower 20% of scores in the PDD-NOS group and the upper 20% of scores in the children with mood and anxiety disorders. Results of a latent class analysis and principle components analysis on the epidemiological sample of school children revealed differences in severity, but not in patterns of scores, suggesting a continuously distributed variable (Constantino et al., 2000).

Strong correlations have been reported between the ADI-R algorithm scores and SRS scores, both based on parent report (Constantino et al., 2004). Principal Component Analysis resulted in single factor explaining 35% of the variance (Constantino et al., 2004). At this point, the SRS is best used as a measure of severity of difficulties in social reciprocity, including odd behaviors. It has been used in genetic studies of ASDs (Constantino & Todd, 2000). The SRS does not take long to administer and demonstrates good reliability. Given that there is overlap between scores in ASD and non-ASD psychiatric populations, its primary use is for measuring symptom severity and response to treatment.

Pervasive Developmental Disorders Rating Scale

The Pervasive Developmental Disorders Rating Scale (PDDRS) is a revision of an earlier scale developed (Eaves, 1990; Eaves & Hooper, 1987), and includes 51 items across three subscales (arousal, affect, and cognition), based on the *DSM-III-R*. Each behavior is rated on a 5-point Likert scale. The author suggests that both the total score and the arousal factor score meet the cutoff of one standard deviation below the mean (standard score > 85), to classify a child as PDD.

The internal consistency, test-retest reliability and inter-rater reliability of the measure were evaluated. Internal consistency was good, resulting in reliability coefficients between .79 and .90 for the scales and .92 on the total score. Test-retest and inter-rater reliability were strong when based on an initial sample in which rating pairs were collected over a mean of 8.33 months, with correlation coefficients between .87 and .91. In a second sample, inter-rater and test-retest reliability were evaluated based on ratings completed by two different respondents over 14 months. Reliability was lower in this situation ranging from .44 to .53 (Eaves, Campbell, & Chambers, 2000).

Convergent and discriminant validity of the instrument were measured by comparing scores on the PDDRS with scores on the ABC and evaluating the sensitivity and specificity of the instrument. Partial correlations, with chronological age as the control variable, were run on the ABC scales and the PDDRS factors. All correlations were significantly different from zero with the exception of PDDRS Cognition and ABC Relating and PDDRS Cognition and ABC Body and Object Use, for which correlations ranged from .32 to .81. The mean score on each scale was significantly higher in the autistic group than a nonautistic group that included nonautism ASDs as well as moderate to severe mental retardation and Williams Syndrome. Using the recommended cut-off score, sensitivity and specificity were 88%. The ABC and the PDDRS scores were consistent in classifying children with autism in 85% of the sample (Eaves et al., 2000). In the validity studies, no standard diagnostic procedure was used to define the sample. Thus, the authors suggest that the instruments be used for screening rather than for diagnostic purposes (Eaves et al., 2000). Because the control group included children with ASDs and there was not a standardized procedure for establishing diagnosis, it is possible that the instrument may miss some children with autism, given that it "screens out" children with related ASDs.

Children's Social Behavior Questionnaire

The Children's Social Behavior Questionnaire (CSBQ; Luteijn, Luteijn, Jackson, Volkmar, & Minderaa, 2000) covers areas associated with

ASDs and was designed to be completed by parents or caregivers of children between the ages of 4 and 18 years. It includes 96 items, 66 of which fall into five factors: Acting-Out, Social Contact Problems, Social Insight Problems, Anxious/Rigid, and Stereotypical (Luteijn et al., 2000). Each item focuses on recent behavior (over the past 2 months) and is rated from zero ("does not describe the child") to two ("clearly applies to the child").

Internal consistency, inter-rater reliability and test-retest reliability were all evaluated for the questionnaire. Internal consistency was fair to excellent with Cronbach's alphas ranging from .76 on the Stereotypical scale to .92 on the Acting-Out scale. Inter-rater reliability between parents was good to excellent, with intra-class correlations ranging from .64 for the Anxious/Rigid scale to .85 for the Social Contacts scale. Test-retest reliability was also good to excellent for most scales, with intra-class correlations ranging from .62 on the Social Insight Problems scale to .90 on the Total, with the exception of the Stereotypical scale, which had a low intra-class correlation of .32 (Luteijn et al., 2000). Convergent and discriminant validity of the scales were measured by comparing scores on the CSBQ with scores on the Children's Behavior Checklist (CBCL; Achenbach, 1981) and the ABC (Krug et al., 1980a) and by comparing mean scores on the measure between diagnostic groups (Luteijn et al., 2000). The scales of the CSBQ were highly correlated with the scales of both the ABC and the CBCL. Three scales of the CSBQ were significantly correlated (.31 to .46) with scores from a checklist based on the *DSM-IV*, completed by a clinician. The exceptions were Acting-Out and Anxious/Rigid, indicating that these two scales were less specific to difficulties associated with an ASD. A discriminant function analysis revealed that 50% of children in the original five groups (PDD-NOS, high-functioning autistic children, attention deficit hyperactivity disorder, clinical control group, mentally retarded children, normal control group) could be correctly classified on the basis of the four discriminant functions: (1) General psychopathology, (2) Withdrawn behaviors, (3) Negative correlation with Social Insight Problems and a positive correlation with Anxious/Rigid, and (4) A strong relationship with

Stereotypical Behaviors and Anxious/Rigid Behaviors (Luteijn et al., 2000).

The authors suggest that the instrument may offer important contributions to research and clinical work, particularly because it revealed different patterns of scores in children with autism and children with PDD-NOS. Specifically, children with PDD-NOS scored higher on the Acting-Out scale than children with autism (Luteijn et al., 2000). One limitation is that the diagnostic groups were determined based on clinical diagnosis alone, rather than with standardized measures. In addition, the correlations of the CSBQ scales with the *DSM-IV* checklist were not high, although they were significant. At this point, the CSBQ remains in the early stages. Further investigation will be important in determining its research and clinical utility.

Achenbach System of Empirically Based Assessment

The Achenbach System of Empirically Based Assessment, Preschool Forms and Profiles (Achenbach & Rescorla, 2000) includes the CBCL for ages 1 year to 5 years, the Language Development Survey (LDS), and the Caregiver-Teacher report form (CTRF).

The CBCL is a questionnaire designed to be completed by parents or caregivers in a home setting, and only requires a fifth-grade reading level. The CBCL scores result in a Total Score, and Internalizing and Externalizing Scale, as well as Syndrome and *DSM* oriented Scales. The *DSM* Oriented Scales include a Pervasive Developmental Disorder Problems Scale that consists of 13 items. Each item is point rated on a 0 to 2 point scale based on behaviors over the past 2 months, with "0" indicating "not true," "1" indicating "sometimes true" or "somewhat true" and "2" indicating "very true" or "often true." Based on raw scores, T-scores can be calculated for each of the *DSM* Oriented Scales. There are cut points for the "borderline range" and the "clinical range." The C-TRF is a teacher rating form designed to be completed by daycare providers or teachers.

While the CBCL is not intended for diagnostic purposes, it is included in this chapter because it includes a Pervasive Developmental Disorders Scale as one of the *DSM* Oriented

Scales. Achenbach and Rescorla (2000) specify that the *DSM* Oriented Scales are not equivalent to a diagnosis, because only behavior over the past 2 months is rated, the behaviors listed do not correspond exactly to diagnostic criteria, and the standard scores are based on age and gender comparisons and the *DSM-IV* is not. However, the scores could be used to identify children with behavior difficulties, and children who have elevated scores (borderline or clinical range) could be referred for further evaluation.

Test-retest reliability was quite high for the Pervasive Developmental Problems scale for both parents ($r = .86$) and teachers ($r = .83$) when the checklist was completed a second time 8 days after the initial rating. Inter-rater reliability on the Pervasive Developmental Problems Scale was moderate between parent to parent ratings ($r = .67$) and teacher to teacher ratings ($r = .67$).

Validity was assessed based on a "clinic referred sample" and a "non-referred sample." As a result, information is not available as to how valid the instrument is for screening specifically for ASD. While the CBCL and C-TRF should not be used as diagnostic instruments, they have potential value as screening tools or research measures of autistic behaviors.

CURRENTLY USED RATING SCALES

Childhood Autism Rating Scale

The Childhood Autism Rating Scale (CARS; Schopler, Reichler, & Renner, 1986) is the strongest, best-documented, and most widely used clinical rating scale for behaviors associated with autism. It has been used in studies all around the world and translated into many languages (Nordin, Gillberg, & Nyden, 1998; Pilowsky et al., 1998; Sponheim, 1996). It consists of 15 items on which children and adults are rated, generally after observation, on a 4-point scale. The scale requires minimal training. Training is available on videotape or in brief workshops. Points are added and a standard cut-off of 30 has been suggested and validated with various samples (Garfin et al., 1988; Schopler et al., 1980). Minor modifications have been suggested in which cut-offs are moved up a few points for very young children

(Lord, 1995) and down for high-functioning adolescents and adults (Mesibov, Schopler, Schaffer, & Michal, 1989).

Most of the information about the CARS is from studies of autistic children who function in the mild to moderate range of mental handicap. Studies of discriminant validity from carefully matched comparison groups are not yet available, though the CARS has been shown to discriminate autistic children from children without autism and some mental handicap (Schopler et al., 1988; Teal & Wiebe, 1986). Convergence between the CARS and the Autism Diagnostic Interview (ADI; Lord, 1995; Sevin et al., 1991; Venter et al., 1992) and correlations between CARS total scores and RLRS total scores (Sevin et al., 1991) were good for autistic children, but less good for young, nonautistic mentally handicapped children (Lord, 1995). Thus, the evidence that the CARS accurately identifies children with autism is stronger than the evidence that it discriminates between children with autism and mental-age matched children with other disorders.

The CARS was created before the introduction of *DSM-IV* and *ICD-10* diagnostic frameworks. It shows good agreement with clinicians' judgments using *DSM-III-R*, though it is somewhat over-inclusive compared to strict application of the criteria (Van Bourgondien, Marcus, & Schopler, 1992). Because, with the exception of the preceding reference, *DSM-III-R* was found to be more inclusive than clinicians' judgments of autism (Hertzog, Snow, New, & Shapiro, 1990; Volkmar, Cicchetti, Bregman, & Cohen, 1992), this finding suggests that the CARS identifies more children as having autism than the currently accepted three-domain diagnostic frameworks of *DSM-IV* (American Psychiatric Association, 1994) and *ICD-10* (World Health Organization, 1992). Children with minimal verbal skills and/or moderate to severe mental handicap may be more likely to fall into the range of autism, in part because items on the CARS rating language skill and mental handicap comprise part of the total score (Pilowsky et al., 1998). For the purposes of screening or determining services, over-inclusiveness of children with clear impairments is not as problematic as over-exclusion (Wing & Gould, 1979). However, implications may be different for research. The CARS cannot be used alone to

make discriminations for complex diagnostic cases in which *DSM-IV* or *ICD-10* criteria are the standard; nevertheless, as discussed earlier, multiple sources are important in any diagnostic decision making and it may provide important information in addition to other sources (Nordin & Gillberg, 1996a, 1996b).

The CARS total score has held up to repeated, careful examinations, as internally consistent (Kurita, Kita, & Miyake, 1992; Sturmey et al., 1992) and reliable across raters (Garfin et al., 1988; Kurita et al., 1992; Sevin et al., 1991). Inter-rater reliability for individual items has been found to be more variable. Some of the scales (e.g., Relating to People, Imitation) have consistently shown high correlations between different raters' scores. Statistics such as kappas, which control for base rates, have not yet been employed (Garfin et al., 1988; Sevin et al., 1991). One of the important contributions of the CARS was the provision of specific anchorpoints for each item in a way that allows the rater to take into account developmental level. The difficulty with this strategy is that how anchorpoints are defined differs across items. Interpretation of scores on individual items, particularly given the inconsistent evidence of reliability at this level, must be carried out with care.

Besides direct observation by a clinician, for which the CARS was designed, it has also been used in chart review, scored directly by parents and teachers, and used as part of a parent interview (Schopler et al., 1988). On the whole, classifications and correlations between raters for total scores have been relatively high across different procedures. Several studies have suggested that clinicians tend to rate behaviors as more severe than do fathers or mothers (Bebko, Konstantareas, & Springer, 1987; Konstantareas & Homatidis, 1989), with other studies finding few differences (Freeman, Perry, & Factor, 1991; Schopler et al., 1988).

A factor analysis of the CARS of 90 children with clinical diagnoses of autism or PDD-NOS based on *DSM-III-R* criteria yielded five factors out of 15 items: Social Communication, Emotional Reactivity, Social Orienting, Cognitive and Behavioral Consistency, and Odd Sensory Exploration. Cognitive and Behavioral Consistency and Emotional Reactivity were significantly correlated with age;

Social Communication was significantly correlated with gender, IQ, and Vineland scores. Factor-based scales distinguished children with autism from those with PDD-NOS. It was suggested that use of these factor scores might increase the sensitivity of the CARS with younger and/or higher functioning individuals within the autistic spectrum (Stella, Mundy, & Tuchman, 1999).

Overall, the CARS is the most widely researched and employed rating scale of autism in the United States. Versions are available in numerous languages other than English. It is a reliable screening instrument for children with autism and mental retardation that can be used with minimal training across a range of situations. Its scores do not correspond to current formal diagnostic frameworks for autism, such as *DSM-IV* and *ICD-10*, and so for research purposes, it may identify a somewhat different population than suggested by those systems.

Autism Behavior Checklist

The Autism Behavior Checklist (ABC) is one component of the Autism Screening Instrument for Educational Planning (ASIEP; Krug et al., 1980b) and the only one that has been evaluated psychometrically. It builds on Rimland's Form E-2, the original Kanner criteria (1943), the Behavior Observation Schedule (Freeman et al., 1978), the BRIAAC (Ruttenberg et al., 1977), and several other sources. It contains 57 items in five areas: sensory, relating, body and object use, language and social interaction, and self-help. It was intended to be completed by teachers as an initial step in educational planning. No special training is required. It has also been used with parents on a retrospective basis for families of high-functioning children (Yirmiya et al., 1994) and on a current basis, yielding somewhat higher scores than with teachers (Volkmar et al., 1988). The rater completes dichotomous ratings, which are weighted according to the authors' data and yield a total score. Ranges, on the basis of a very large, but unspecified sample, are provided for a high probability of autism (≥ 68), low probability of autism (under 53), and mixed. Several investigators have reported that the suggested cut-offs are too high, and result in a high proportion of false negatives (Miranda-Linne et al., 1997; Volkmar

et al., 1988; Wadden, Bryson, & Rodger, 1991). More recently, Krug, Arick, and Almond (1993) recommended using a cut-off of greater than 53 for classifying a child as probably autistic. When using this lower cut-off, Eaves et al. (2000) found that overall classification accuracy was 80%, specificity (correct negatives) was 91% and sensitivity (correct positives) was 77%. Norms and standard profiles are provided for samples of autistic, typical, deaf, and blind students.

Initial estimates for inter-rater reliability were high, though based on small samples and not controlling for chance (Krug et al., 1980b). Later estimates have been less high (Volkmar et al., 1988). Discriminant validity has been variable, in part depending on whether investigators generated discriminant functions from data within their group or used the cut-offs suggested by the authors. In the latter case, there was considerable overlap between autistic and mentally handicapped populations (Volkmar et al., 1988). In the former case, diagnostic differentiation was, not surprisingly, better (Nordin & Gillberg, 1996b; Wadden et al., 1991). Current scores on the ABC did not meet criteria for most of a group of verbal adolescents with autism, but retrospective accounts did (Yirmiya et al., 1994). Differences in studies may also be related to the use of a somewhat broader definition of autism, in which case the ABC becomes more accurate in diagnosing autism, and inclusion of subjects with Down syndrome, which may decrease the false positive rate (Wadden et al., 1991).

Internal consistency for the total scale is good. Various investigations have yielded different results in terms of the internal consistency and intercorrelations of the five areas; both chronological and mental age may account for much of the variance. Subscales of relating and object/body use were the strongest in one study in terms of inter-item correlations and lack of rogue items (Sturmeiy et al., 1992). Several investigators have suggested that discriminant validity may be equally good using fewer items (Volkmar et al. 1994; Wadden et al., 1991).

Convergent validity between the ABC and other instruments has been measured for the CARS and the RLRS and found to be poor, suggesting that the ABC's usefulness as an

independent diagnostic instrument may be limited, particularly since it was constructed before current theoretical frameworks for autism were proposed (Nordin & Gillberg, 1996b). For verbal autistic adolescents, retrospective parent ratings on the ABC about their children's behavior between 3 and 5 years, related to whether children were considered to have "residual" autism or not, but diagnosis did not correspond to the cut-offs suggested by the authors of the scale (Yirmiya et al., 1994).

The ABC emphasizes autistic symptomatology rather than prosocial behaviors and so is quite different than several of the other instruments, for example, the ADI-R. Because of its emphasis on observable features associated with, but not limited to autism, the ABC may be helpful in documenting change. This would be particularly true for changes in the presence of abnormal behaviors. Unlike several other autism scales that showed more consistent convergent validity with each other, the ABC is correlated with the American Association of Mental Deficiency (AAMD) Adaptive Behavior Scale-School Version (Sevin et al., 1991). The ABC alone cannot be considered a strong diagnostic instrument because of its limited relationship to current diagnostic frameworks. As it stands, it is of limited value as a screening instrument because of variable sensitivity. However, the ABC may be useful in documenting response to treatment and educational programming.

Revised Behavior Summarized Evaluation

The Revised Behavior Summarized Evaluation (BSE-R) is composed of items from two overlapping instruments, the Behavioral Summarized Evaluation scale (BSE) and the Infant Behavioral Summarized Evaluation scale (IBSE; Barthelemy et al., 1997) and is primarily designed to document behavioral symptoms associated with autism as they relate to neurophysiological measures. New items have been added concerning nonverbal communication, emotion, and perception, as well as intention and imitation. These scales are available in French and have been used in many basic research investigations of children with autism in France (for example, see Zakian, Malvy, Desombre, Roux, & Lenoir, 2000). There are

20 items in the BSE selected from 19 items from the autism factor in the IBSE, the form for children under 4 years of age (Adrien et al., 1992) and 20 in the original BSE (Barthelemy et al., 1990). Items are scored on a five-point scale administered by trained raters, on the basis of direct or videotaped observation, discussion of history, and access to information from multiple sources. With trained raters, most individual items have shown very good inter-rater reliability. Inter-rater reliability for total scores has been excellent, though ratings were not typically based on independently acquired information.

Factor analyses have shown loadings within one primary Interaction Disorder factor, accounting for 38% of the variance and a Modulation factor, accounting for 10%. Results from previous versions indicated adequate internal consistency (Adrien et al., 1992; Barthelemy et al., 1990). The Interaction factor was not correlated with age but was highly negatively correlated with IQ ($r = -.59$). Discriminant function analyses accurately grouped 80% to 85% of autistic and mentally handicapped children using the IBSE (Adrien et al., 1992). Interaction Disorder factor scores were correlated with expert ratings of severity of autism (Barthelemy et al., 1990, 1997). A cut-off score of 27 on the Interaction Disorder factor on the BSE-R yielded a sensitivity of .74 and a specificity of .71 (Barthelemy et al., 1997). Convergent validity with other measures except the Rimland E2 is not yet published. There is some suggestion that the BSE-R may be particularly helpful in measuring response to treatment (Boiron, Barthelemy, Adrien, Martineau, & Lelord, 1992) and in neurophysiological studies (Barthelemy et al., 1997).

The Gilliam Autism Rating Scale

The Gilliam Autism Rating Scale (GARS; Gilliam, 1995) is a parent-completed surveillance questionnaire, designed to indicate the probability that a child has autism. It is intended for individuals between 3 and 22 years of age. The questionnaire consists of 56 items across four subscales: Social Interaction, Communication, Stereotyped Behaviors, and Developmental Disturbances. The first three subscales listed are based on a child's current

behavior, and the final scale is based on a child's developmental history. Each item is rated on a four-point scale, from "Never Observed" to "Frequently Observed." Item scores are totaled for each scale and correspond to a standard score with a mean of 10 and a standard deviation of 3. Typically, all scales of the GARS are completed. However, if a child is nonverbal and/or the parent does not have knowledge of the child's early history, the Communication or Developmental History scales may be omitted. A standard score or Autism Quotient can be based on 4, 3, or 2 scales of the GARS. An Autism Quotient is derived by summing relevant scale scores, yielding a standard score with a mean of 100 and a standard deviation of 15. The Autism Quotient is divided into seven ordinal categories, ranging from a "low probability" that a child has autism to a "high probability" that a child has autism.

Internal consistency of the items on the scale using Cronbach's alpha yielded coefficient alphas ranging from .88 to .96 (Gilliam, 1995). Correlations among individual GARS scales rating current behaviors are relatively high. The Developmental Disturbances scale was not significantly correlated with any of the other scales, although it was weakly, but significantly, correlated with the Autism Quotient, $r = .34$ (South et al., 2002). Test-retest reliability on a small sample of 11 children, using three of the scales (excluding Developmental History) revealed correlations between totals ranging from $r = .81$ (Communication) to $r = .88$ (Autism Quotient). Neither item agreement nor classification agreement across time or rater were reported. Finally, inter-rater reliability was evaluated for the various pairs of raters (parent-parent, teacher-teacher, and parent-teacher). Teacher-to-teacher and teacher-to-parent inter-rater reliability estimates were all strong, ranging from .85 to .99. Ratings were weakest for the Parent-to-Parent ratings with reliability ranging from .55 to .85 (Gilliam, 1995).

The initial reference sample for the GARS consisted of data collected for 1,092 children, adolescents, and adults (Gilliam, 1995). Although a parent or professional rater reported each individual's diagnosis, an independent professional did not verify it. Thus, there was

no "gold standard" for diagnosis. Concurrent validity was evaluated by correlating standard scores on the GARS with scores on the ABC; all correlations were large and significant. Discriminant validity was evaluated by determining how well the measure discriminated between groups that were diagnosed with autism as compared to those who were not. Significant differences were found between the means of those diagnosed with autism versus those who were not. Using the Autism Quotient, 90% of the subjects were classified accurately.

A later independent study of the validity of the GARS was based on a sample of 119 individuals with autism, all of whom had extensive diagnostic evaluations, including the ADOS and the ADI-R, by experts in ASDs (South et al., 2002). The validity data from this study was disappointing, with the GARS receiving a sensitivity of .48 compared to "gold standard" diagnoses, indicating that 52% of children with autism (based on the ADI, ADOS, and clinical impression) were missed by this instrument (South et al., 2002). Convergent validity was also investigated by comparing the GARS scores to scores on the ADOS and the ADI-R (South et al., 2002). There were no significant correlations between any of the GARS scales and the ADOS. Small but significant correlations were reported between the ADI-R Social Interactions score and the GARS Social Interaction Scale ($r = .26$), Stereotyped Behaviors Scale ($r = .21$), and the Autism Quotient ($r = .23$).

Given the current information about its validity and the high rates of false negatives, the GARS cannot yet be used in isolation as a diagnostic tool. This is particularly concerning because, although most research projects have used the instrument in conjunction with other instruments (Asano et al., 2001; Owley et al., 2001), some studies have employed the GARS for diagnostic purposes (Schreck & Mulick, 2000). The use of this instrument may be even more problematic in clinical settings where the professionals know less about autism. In this context, many children with autism could be missed and not referred for appropriate services. The author intends to revise the instrument and has proposed using a lower cut-off score (South et al., 2002).

DIAGNOSTIC INTERVIEWS

Autism Diagnostic Interview-Revised

The Autism Diagnostic Interview-Revised (ADI-R) is a semi-structured, investigator-based interview for caregivers of children and adults for whom autism or pervasive developmental disorders is a possible diagnosis. Originally developed as a research diagnostic instrument (ADI; Le Couteur et al., 1989), the ADI-R has been modified to be appropriate for a broader age range of children than the original ADI (Lord et al., 1994). It is linked specifically to *ICD-10* and *DSM-IV* criteria. A revised shortened version is now available, consisting of about 93 items. The most recent version takes about 2 hours for an experienced interviewer to administer (Le Couteur, Lord, & Rutter, 2003). Researchers are required to participate in training workshops and to establish reliability with investigators from other centers. Clinicians are encouraged to use video training materials, and may use the instrument without intensive training within the ethical guidelines for test use in their professions. Nonetheless, administering the ADI-R requires general experience in both interviewing and working with individuals with autism to be effective. The ADI-R has been translated into 11 languages and the ADI and the ADI-R are cited as the “gold standard” for diagnosis in many countries.

Psychometric data for the ADI and ADI-R have been carefully acquired with attention to matching across samples and to maintaining as much “blindness” as possible for raters, but is based on very small samples (Rutter et al., 2003). This limitation is compensated for slightly by independent psychometric data published by other major research centers that have used the ADI or ADI-R as a diagnostic instrument (Constantino et al., 2004; Cuccaro, Shao, Grubber, et al., 2003; deBildt et al., 2004; Kolevzon et al., 2004; Saemundsen, Magnússon, Smári, & Sigurdardóttir, 2003). Inter-rater reliability has been good to excellent for individual items and excellent for domain scores, including those for each of the three subscales: social reciprocity, communication, and restricted, repetitive behaviors that correspond to the *DSM IV/ICD-10* do-

main (see Chakrabarti & Fombonne, 2001; Rutter et al., 2003). Test-retest reliability, on a very small sample, was also good (Lord et al., 1994). Change over time is reflected in items that include whether the behavior “ever” occurred and items that focus on “current” manifestations. On the whole, however, the ADI-R is not intended to measure change. There has been a deliberate attempt to include items that will reflect autism of varying levels of severity and at varying points in development.

Internal consistency is excellent within the three domains. Differentiation between autistic and mentally handicapped children and adults is excellent, with the restriction that the instrument tends to be over-inclusive for individuals with mental ages of less than 18 months (Lord et al., 1993) and with severely to profoundly retarded individuals (Nordin & Gillberg, 1998). One study found that the ADI was slightly under-inclusive with very verbal children with autism or pervasive developmental disorders (Yirmiya et al., 1994); another study reported that it was over-inclusive (Mahoney et al., 1998). Convergent validity with the CARS was excellent after age 3 (Lord, 1995; Pilowsky et al., 1998); convergent validity with the Autism Diagnostic Observation Schedule (see below) has also been good for most samples (Hepburn et al., 2003; Lord, Risi, et al., 2000; Lord et al., 1989). The exception was a recent study of Bishop and Norbury (2002) in which children with language impairments, in some cases also with ASD, were given the ADI-R or SCQ, the ADOS, and the Communication Competence Checklist (CCC). ADI-R and SCQ classifications were comparable to school classifications but not with ADOS scores or scores on the CCC, which were similar to each other. It was not clear if this was related to specific difficulties using the ASD instruments in a relatively narrowly defined verbal sample, administration of a single module of the ADOS regardless of language level in some cases (all children were given Module 3) or differences in parent report and school classification systems and direct observations.

Because of the widespread use of the ADI-R in defining samples, there has been a recent surge of interest in how to use the ADI-R for a variety of other purposes beyond classification,

including quantifying severity (Lord, Leventhal, & Cook, 2001; Spiker, Lotspeich, Dimiceli, Myers, & Risch, 2002; Szatmari et al., 2002; Volkmar & Lord, 1998), describing individual differences (Alarcón et al., 2002; Cuccaro, Shao, Bass, et al., 2003; Tanguay, Robertson, & Derrick, 1998) and creating more homogeneous subsets of participants for genetic analyses (Buxbaum et al., 2001; Freitag, 2002; Shao, Raiford, et al., 2002; Shao, Wolpert, et al., 2002; Tadevosyan-Leyfer et al., 2003). These studies have used a wide range of analytic techniques, sometimes related to different purposes, and have been carried out on a wide range of items (e.g., sometimes all ADI-R item scores are included; sometimes selected items, sometimes domain scores). Studies have varied considerably whether age, IQ or verbal level were controlled. It is clear that, depending on the ranges studied, all three of these features can affect ADI/ADI-R scores (Cox et al., 1999; Cuccaro, Shao, Bass, et al., 2003; Spiker et al., 1994).

Overall, no differences were found in domain scores for multiplex families compared to singletons (Cuccaro, Shao, Bass, et al., 2003). Factor analyses of domain scores and Vineland adaptive behavior scores (Sparrow, Balla, & Cicchetti, 1984) in two separate samples yielded a symptom number factor and a separate factor for level of functioning, determined by the Vineland adaptive behavior scores (Szatmari et al., 2002). Only the ADI-R domain of nonverbal communication showed any evidence of concordance within multiplex families (MacLean et al., 1999), a relationship also found by another research group (Freitag et al., 2002). In another sample, heritability was supported for a continuous severity gradient composed of ADI-R scores, verbal—nonverbal status and nonverbal IQ (Spiker et al., 2002). Several other studies that have measured concordance within twin pairs (Le Couteur et al., 1996) and families (Spiker et al., 1994) found contradictory results with little concordance on any dimension for monozygotic twins, but concordance for ADI-R repetitive scores found in families (Freitag et al., 2002; Spiker et al., 1994).

Several groups of genetics researchers have produced increased homogeneity and more significant results by subsetting groups by individ-

ual items within the ADI-R repetitive domain (Alarcón et al., 2002) or by the entire domain score (Silverman et al., 2002). Other studies have found that only particular combinations of items (e.g., insistence on sameness, compulsions) yielded similar results (Shao et al., 2003) for other genetic regions. Though potential genetic significance of repetitive behaviors emerges across papers, in most cases, studies have not replicated each other, nor have age, IQ, or verbal status been controlled consistently.

The other way in which the ADI-R has been used within genetic studies has been to produce subsets based on language delay (based on measures of age of first word or age of first phrase) or on current language level. Concordance for current verbal ability has been shown in some cases (Freitag et al., 2002; MacLean et al., 1999; Spiker et al., 2002), but on the whole, only delay in either first single words (Alarcón et al., 2002) or first phrases (Bradford et al., 2001; Buxbaum et al., 2001; Shao, Wolpert, et al., 2002) or delay accompanied by the presence of a language-delayed relative (Folstein & Mankoski, 2000) increased the significance of specific regions.

Several factor analyses and principal component analyses have been carried out, primarily with data from earlier versions of the ADI. In one study, factors emerged that reflected three aspects of social communication: Affective Reciprocity, Theory of Mind, and Joint Attention (Tanguay et al., 1998). In another study (Lord, 1990), social and communication items both loaded on two factors; in this case, the factors seemed to reflect initiations versus social responsiveness. In a recent study, six factors emerged, that together accounted for about 40% of the variance (Tadevosyan-Leyfer et al., 2003). These factors consisted of items scored for both current functioning and “ever”/most abnormal 4 to 5, that were present in both the early ADI and the ADI-R so they represent a particular subset of questions. Factors were validated in another sample using additional psychometric measures. Constantino and Todd (2003) recently reported a factor analysis of the ADI-R, with a different pattern.

It seems very likely that items within the ADI-R can be combined in more fruitful ways than the present algorithm domain scores. One consistent finding across these studies is the

overlap between “communication” and “social” items, suggesting that they are not separate domains of skill (Lord, 1996; Tadevosyan-Leyfer et al., 2003; Tanguay et al., 1998). Several factors with different organizations of repetitive behaviors have also been proposed. To date, however, factors in this area and across other domains have differed considerably across investigations. The development of a more stable measure or measures of repetitive interests and behaviors will be an important contribution to better understanding of phenotypes in ASDs. Larger samples, including individuals without autism, will be necessary in order to control effects of age, verbal status and IQ. Replication across sites and samples will be crucial in determining the factors of greatest interest or usefulness.

ADI scores have also been shown to be related to ABC scores, given by history, for a group of high-functioning children (Yirmiya et al., 1994). Because of its clear link to *DSM-IV* and *ICD-10* and its multidimensional approach, the ADI-R offers the potential of providing empirical information and diagnostic guidance about other PDDs besides autism. However, cut-offs for nonautism pervasive developmental disorders are not yet available. Several investigations have proposed various cut offs, including one or two points below autism thresholds (over all the domains), but none have yet been empirically validated (Cox et al., 1999; Dawson et al., 2002).

The Diagnostic Interview for Social and Communication Disorders

The Diagnostic Interview for Social and Communication Disorders (DISCO: Wing et al., 2002) is a standardized, semi-structured interview, now in its ninth revision. It is based on the Handicaps, Behaviors and Skills schedule (HBS) (Wing & Gould, 1978, 1979). In 1990, a clinical need emerged for an instrument that extended beyond the school age years, into adulthood. At this time, the first version of the DISCO was developed to assess the pattern of development in individuals with ASDs and their individual needs (Wing et al., 2002). The primary purpose of the DISCO is not to provide a diagnostic classification. Rather, the instrument was designed to obtain information

on behaviors relevant to autism for the purpose of assisting clinicians in determining a child’s development in different areas as well as his individual needs (Leekam, Libby, Wing, Gould, & Taylor, 2002). It is based on the concept of a spectrum of disorders rather than categorical diagnoses.

The DISCO is an investigator-based interview in which the interviewer asks questions designed to elicit descriptions of behavior and makes coding decisions based on the information provided. The coding of the items can be based on information obtained during the interview as well as through other information, such as direct observation. The DISCO includes items covering behavioral manifestations of the deficits associated with ASDs, including social interaction, communication, imagination, and repetitive activities. In addition, it includes items designed to assess developmental levels in a variety of domains. Many of these items are based on the Vineland Adaptive Behavior Scales (Sparrow et al., 1984). There is also a section on atypical behaviors that are not specific to autism. These include unusual responses to sensory stimuli, difficulties in attention and activity level, challenging behaviors, and other psychiatric disorders. Items relating to developmental delay are rated on a 3-point scale, as “delay,” “minor delay,” or “no problem.” An actual age is coded for some of the developmental items. Atypical behaviors receive codes for “current” and “ever” and are rated as “severe,” “minor,” or “not present.”

The reliability of the DISCO 9 was evaluated based on a sample of 82 children with diagnoses of ASDs, learning disability, or no diagnosis (typically developing) between the ages of 3 and 11 years of age (Wing et al., 2002). Inter-rater reliability was measured, comparing two interviewers/coders, using Kappa’s alpha for items with two or three codes and by intraclass correlations (ICC) for items with four or more codes. Agreement was high (k or $ICC > .75$) for 85% of all ratings for both preschool age and school age children. Within the Developmental Skills area, the lowest agreements (.67 to .80) were for items that were not part of the diagnostic algorithm (e.g., reading, drawing). Of greater concern, was the low agreement (with kappas $< .40$) on some of

the social interaction items and for many of the repetitive routine items, which are part of the diagnostic algorithm. Inter-rater reliability was higher for the “ever” items than for the “current” items. Based on this information, the authors plan to make some changes designed to improve reliability, which will be included in the DISCO 10.

While the DISCO was designed for clinical purposes, provisional algorithms have been written for research purposes. Recently, two diagnostic algorithms for the DISCO 9 were developed and investigated (Leekam et al., 2002). One of the algorithms was based on criteria for autistic disorder in the *ICD-10* (World Health Organization, 1992) and the other was based on the criteria for autistic spectrum disorder as defined by Wing and Gould (1979). When comparing clinical diagnosis to algorithm diagnoses for a sample of children with language disorders, learning disability, and autistic disorder, both algorithms were significantly related to a diagnosis of autistic disorder or nonautistic disorder. However, discrepancies were also found, primarily for the clinical nonautistic group using the *ICD-10* algorithm such that 10 children with clinical diagnoses of a language disorder or learning disability met *ICD-10* algorithm criteria for autistic disorder. Four children with a learning disability diagnosis met both *ICD-10* and Wing and Gould algorithm criteria for autistic disorder, while none with a language disorder met criteria using both algorithms. DISCO 9 algorithms were also generated for Gillberg’s diagnostic criteria for Asperger’s Disorder and *ICD-10* criteria for Asperger’s Disorder. Of the 200 children included in the study, all of whom met *ICD-10* criteria for autism or atypical autism, only 3 (1%) met criteria for Asperger’s Disorder based on the *DSM-IV* algorithm and 91 (45%) met criteria based on Gillberg’s algorithm criteria (Leekam, Libby, et al., 2000).

The DISCO was primarily designed for clinical purposes, particularly for assisting in generating recommendations for individuals and adults with autistic spectrum disorders. The authors are revising the instrument to improve inter-rater reliability and to generate diagnostic algorithms that can be used for research purposes.

DIRECT OBSERVATION SCALES

Autism Diagnostic Observation Schedule

The Autism Diagnostic Observation Schedule (ADOS) is a standardized protocol for the observation of social and communicative behavior of children for whom a diagnosis of autism or ASDs is in question (Lord, Rutter, DiLavore, & Risi, 1999; Lord, Risi, et al., 2000). The original ADOS was developed in order to be used with children who had fluent phrase speech; the Pre-Linguistic Autism Diagnostic Observation Schedule (PL-ADOS) was intended for preschool children with little or no expressive language (DiLavore et al., 1995; Lord et al., 1989). Recently they have been combined and extended within a single instrument, the Autism Diagnostic Observation Schedule (Lord, Rutter, DiLavore, & Risi, 1999), formerly called the ADOS-G, with the PL-ADOS comprising most of Module 1, the original ADOS comprising most of Module 3, and the addition of new modules for children with some language but not fluent spontaneous speech (Module 2) and for high-functioning adolescents and adults (Module 4). The new ADOS thus provides the same information as the original ADOS and PL-ADOS for individuals ranging in age and development from nonverbal toddlers to verbally fluent adults of average or higher intelligence.

The ADOS and PL-ADOS were originally developed as companion instruments for the ADI. Their purpose is to provide a series of structured and semi-structured “presses” for social interaction, communication, and play that can be coded immediately following administration (although often videotapes are made as well). They are scored in the context of a diagnostic algorithm for autism. The rationale is that context can have very significant effects on social-communicative behaviors. Consequently, it is important to standardize contexts as well as judgments in any diagnostic observation of these behaviors. Both instruments can be administered by a trained examiner in about 30 to 45 minutes. Training and establishment of reliability with another center is required for research, but not for clinical use. A substantial amount of experience, skill, and practice in working with individuals with

autism or PDD is necessary to use either instrument effectively.

Inter-rater reliability is very good for items and excellent for totals. Internal consistency within domains of social-communication and restricted-repetitive behaviors is excellent (Lord, Risi, et al., 2000); test-retest reliability is adequate. Discriminant validity is excellent for diagnostic algorithms using social-communication scores. In the normative data, within each module, social and communication scores were relatively independent of absolute expressive language level. However, recent studies have found relatively strong effects of level of verbal impairment (e.g., verbal IQ), communication scores, and for social domain scores, particularly with preschool children (Munson, Dawson, Lord, Rogers, & Sigman, in press).

The instruments were expanded into the four modules that comprise the ADOS because of varying problems of sensitivity and specificity by age and language level. Diagnostic algorithms for the PL-ADOS were under-inclusive for children with phrase speech with about 80% accuracy overall for autistic and/or nonautistic mentally handicapped 3- and 4-year-olds. For the original ADOS, the diagnostic algorithm was over-inclusive for children with mental handicap and difficult behaviors and was under-inclusive for very verbal adolescents, with about 87% accuracy comparing autistic to mentally handicapped and behavior-disordered, language impaired children. The design of four different modules has increased the diagnostic accuracy of the ADOS considerably, but nevertheless, it remains over-inclusive with very young (under 30 months), mentally retarded children (Hepburn et al., 2003), and under-inclusive with very mild, verbal adolescents and adults with autistic spectrum disorders (International Molecular Genetics Study of Autism Consortium, 1998; Lord, Risi, et al., 2000).

A recent factor analysis of the original ADOS yielded three factors that accounted for 72% of the variance (Robertson, Tanguay, L'Ecuyer, Sims, & Waltrip, 1999) in a sample of verbally fluent children with ASD. These factors, similar to those found in an analysis by the same authors of ADI data (Tanguay et al., 1998), were Joint Attention, Affective Reciprocity, and Theory of Mind. Theory of Mind scores on the two instruments were correlated

$r = .31$, but scores on the other factors were not. Together with findings from the related analysis of the ADI-R, these results highlight the importance of considering social development and communication together in the use of these diagnostic instruments.

Like the ADI, the ADOS was not originally intended to measure change, although it may be possible to use the standard behavior samples provided by the ADOS in conjunction with other coding systems as a measure of response to treatment (Owley et al., 2001). As is the case with the ADI, it is hoped that multidimensional scoring of the ADOS may allow for better quantification of nonautism pervasive developmental disorders, most notably PDD-NOS and Asperger's disorder. Clinically, the ADOS is particularly helpful in providing information concerning social and communicative functioning, which has been collected in a positive but standard context, to parents, therapists, and teachers.

The Psychoeducational Profile-Revised

The Psychoeducational Profile (PEP; Schopler & Reichler, 1979) is a developmental and diagnostic assessment instrument designed specifically for assessing children with ASDs. It was revised in 1990 (PEP-R; Schopler, Reichler, Bashford, Lansing, & Marcus, 1990), is currently under revision once again, and will soon be available as the Psychoeducational Profile-Third Edition (PEP-3). The instrument has been translated into several different languages. The PEP and PEP-R are most appropriate for use with children between the chronological ages of about 3 and 7 years. The normative sample included 420 children between 1 year and 7 years of age. Much of the available published information covers the PEP as well as the PEP-R. Because this is a chapter on diagnosis, only the pathology scales will be reviewed. The pathology scales of the PEP-R are designed to rate the severity of the characteristics of autism in the following areas: Response to Materials (8 items), Language (11 items), Affect & Development of Relationships (12 items), and Sensory Modalities (12 items). On these scales, pathology is rated as "absent," "mild" or "severe."

Both convergent and discriminant validity of the original PEP's pathology section have been evaluated. Schopler and Reichler (1979) reported a high correlation ($r = .80$) between pathology scores on the Childhood Autism Rating Scales (CARS) and pathology scores on the PEP. When comparing children with autism with children without autism, children with autism exhibited higher pathology scores on the PEP (Lam & Rao, 1993). Internal consistency of the diagnostic section of the PEP-R pathology subscales has been reported to be good, with Cronbach's alphas between .84 and .97 (Steerneman, Muris, Merckelbach, & Willems, 1997). Very little information is available on the inter-rater reliability of the pathology subscales of the PEP or the PEP-R. The one study that investigated inter-rater reliability reported a mean kappa score of .69 (Muris, Steerneman, & Ratering, 1997), which is considered adequate.

Much of the information available on the PEP is based on the original version of the measure. Many of the studies are small scale studies conducted on translations of the instrument. The developmental scores of the PEP-R have been used in outcome studies (Ozonoff & Cathcart, 1998; Panerai, Ferrante, & Caputo, 1997). In research, the pathology subscales are used much less frequently than the developmental subscales, which are most often administered to establish the developmental levels of lower functioning children with ASDs or to evaluate treatment outcomes. The PEP is frequently used in conjunction with the CARS in research studies measuring both diagnostic classification and developmental level.

The Adolescent and Adult Psychoeducational Profile

The Adolescent and Adult Psychoeducational Profile (AAPEP; Mesibov, Schopler, & Caison, 1989) is an extension of the PEP and was also developed by Division TEACCH. Like the PEP, the AAPEP is designed to assess individuals with ASDs for the purpose of developing individualized treatment goals and recommendations. The AAPEP is a criterion-referenced test and targets individuals over 12 years

of age with moderate to severe mental retardation. As a result, the targeted areas focus on concerns that often appear as adulthood approaches and include matters such as semi-independent functioning and psychopathology in the community.

The AAPEP incorporates three separate scales: a direct observation scale and two interview sections (a home scale and school/work scale). Each scale includes six functioning areas: vocational skills, independent functioning, leisure skills, vocational behavior, functional communication, and interpersonal behavior. Little information is available on the validity and reliability of this instrument. Inter-rater reliability was evaluated by calculating the percent agreement between two independent raters and was determined to be sufficient (with the exception of Interpersonal Behaviors on the Direct Observation scale; $r = .68$), ranging from $r = .74$ to $r = .95$ (Mesibov, Schopler, & Caison, 1989; Mesibov, Schopler, Schaffer, & Landrus, 1988). There has been little research using the AAPEP; however, one study evaluated progress in adults with ASDs who were living in a group home setting (Persson, 2000).

There are few scales available for measuring functional behaviors and skills in adults with autism. The AAPEP is not intended for diagnostic purposes and focuses primarily on the assessment of skills required for independent living. The best application of the AAPEP is for identifying target areas for intervention or skill building.

RELATED DIAGNOSTIC AND BEHAVIORAL ASSESSMENT INSTRUMENTS

The Communication and Symbolic Behavior Scales, Developmental Profile

The Communication and Symbolic Behavior Scales, Developmental Profile (CSBS DP) is a standardized instrument designed for screening and evaluating communication and symbolic abilities in young children between the ages of 6 and 24 months (Wetherby & Prizant, 2002). The published version is based on an earlier version designed specifically for research pur-

poses (Wetherby & Prizant, 1993). There are three separate parts of the CSBS DP, including a screening instrument (CSBS DP Infant and Toddler Checklist) and two follow up assessment tools: a parent questionnaire (CSBS DP Caregiver Questionnaire) and a direct observation section (CSBS DP Behavior Sample). The purposes of the CSBS DP are screening and identifying children at risk for language and developmental delays, not specifically autism, as well as assessment and identification of delays in social communication, expressive language, and symbolic abilities. The CSBS DP also provides an opportunity for documentation of progress over time. The CSBS DP consists of seven cluster areas (Emotion and Eye Gaze, Communication, Gestures, Sounds, Words, Understanding, Object Use) that are included in one of three composites (Social Communication, Expressive Speech & Language, and Symbolic Abilities).

Although the CSBS was not specifically designed to screen or evaluate young children with ASDs, there is evidence that information gathered from the Behavior Sample may have some value in screening for ASDs. Wetherby et al. (in press) compared children with ASDs, children with developmental delays, and children who were typically developing using the Systematic Observation of Red Flags (SORF) for Autism Spectrum Disorders in Young Children (Wetherby et al., in press), which is based on the Behavior Sample of the CSBS. The SORF includes 29 items from both the diagnostic criteria and research on ASDs in young children. It covers five composite areas and includes reciprocal social interaction, unconventional gestures, unconventional sounds and words, repetitive behaviors and restricted interests, and emotional regulation (Wetherby et al., in press). Inter-rater reliability for the SORF was high (89.7% to 100% agreement across children and 83% to 100% across items).

The SORF shows promise as a screening instrument, with sensitivity, specificity, positive and negative predictive values all over 80% (Wetherby et al., in press), based on the Behavior Sample. A discriminant function analysis indicated that when 15 red flags were considered, 100% of the children in the ASD group, 83% of the developmentally delayed group, and 100%

of the typical group were correctly predicted (Wetherby et al., in press). At this point, there are no cut-offs suggested, but children who demonstrate most of the 15 red flags should be referred for further evaluation. More investigation using the SORF and CSBS behavior sample with children with ASDs is warranted as it shows promise as a valuable screening tool.

The Children's Communication Checklist

The Children's Communication Checklist (CCC) was developed by Dorothy Bishop (1998) to assess pragmatic difficulties within the speech and language impaired population. Although there are several standardized tests available for assessing language form, such as syntax and phonology, adequate standardized assessment instruments for assessing pragmatic difficulties are very rare. The CCC is designed to be completed by a professional, such as a teacher or a speech and language therapist, who knows the child well (Bishop, 1998; Bishop & Baird, 2001). It consists of five scales assessing pragmatic aspects of speech: inappropriate initiation, coherence, stereotyped language, use of context, and rapport. In addition, it includes two item sets designed to assess other aspects of speech and language (speech production and syntactic complexity), as well as two item sets intended to assess nonlanguage features of autistic spectrum disorders (social relationships and interests). Each behavior is described, and the rater is asked to indicate if it "definitely applies," "applies somewhat," "does not apply," or if they are "unable to judge." Most of the data available on the CCC is based on children between the ages of 5 and 17 years.

There is debate in the literature as to whether there is a pure group of children with pragmatic difficulties and the extent to which they overlap with children with ASDs (Bishop, 1998; Botting & Conti-Ramsden, 1999). At least a subset of children with pragmatic language difficulties also meet criteria for an autistic spectrum disorder (Botting & Conti-Ramsden, 1999). This has led some individuals to hypothesize that Autistic Spectrum Disorders and Pragmatic Disorder may be related both in symptoms and etiology (Bishop, 1998). Although individuals with receptive-expressive

language disorders generally tend to function better than adults with autism, when adults who were diagnosed with receptive-expressive impairments as children are assessed with the ADI-R and ADOS, there is considerable overlap in adult diagnosis. In one study, 60% of the developmental language disorder group was misclassified as autistic on at least one variable (social functioning, independence, or ritualistic/stereotyped behavior) and 33% of the autistic adults were misclassified as language impaired (Howlin et al., 2000). Other studies have found that when a score of less than 132 on the CCC (lower scores indicating greater pragmatic difficulties) is used as a cut-off, children with autism, semantic-pragmatic language impairments, or semantic-pragmatic language impairments plus autistic characteristics have lower scores than a speech language impaired group (Bishop, 1998). Another study found using this same cut-off, children with autism had lower scores than a learning disabilities group (Botting & Conti-Ramsden, 1999). The purpose of the CCC, however, is not to differentiate children with a language disorder from the general population, but rather to differentiate pragmatic difficulties from other aspects of language disorder within the language impaired population (Bishop, 1998; Bishop & Baird, 2001).

Validity of the instrument was evaluated by comparing scores on the CCC between three different diagnostic categories (Semantic-pragmatic pure—did not have autistic symptoms; Semantic-pragmatic plus—did have autistic symptoms or an autistic disorder, Other speech and language impairment—without pragmatic difficulties or autistic characteristics), based on school system classifications (Bishop, 1998). Based on this study, children with a composite score lower than 132 were more likely to be in the semantic-pragmatic pure or the semantic-pragmatic plus (pragmatic disorder plus some autistic characteristics) groups and those children with scores higher than 132 were more likely to be in the other speech and language impaired group (Bishop, 1998). Interestingly, parent ratings on the CCC relate more clearly to the child's diagnostic status than do ratings by teachers. The authors recommend combining

parent and professional report to obtain the most accurate information.

INSTRUMENTS FOR ASPERGER'S DISORDER

The Asperger's Syndrome (and High-Functioning Autism) Diagnostic Interview

The Asperger's Syndrome (and High-Functioning Autism) Diagnostic Interview (ASDI) was developed as a diagnostic tool specifically tailored for verbally fluent autism and Asperger's Disorder (Gillberg, Gillberg, Rastam, & Wentz, 2001). The interview is based on Gillberg's diagnostic criteria for Asperger's Disorder, and includes 20 items that operationalize six criteria (Social, Interests, Routines, Verbal and Speech, Communication, and Motor). The ASDI is a structured interview that is administered to a person who knows the subject of the interview quite well, and has some knowledge of the subject's childhood. Each question is rated on a three-point scale. The interviewer is instructed to obtain details on actual behaviors to accurately code each item.

Initial reliability studies were conducted on a group of 20 individuals between 6 and 55 years of age. Inter-rater reliability was investigated and results indicated exact agreement for 96% of the ratings (383 out of 400 ratings), resulting in a kappa of .91. Test-retest reliability was also investigated, and complete agreement was achieved for 97% of the ratings (465 out of 480), resulting in a kappa of .92.

Validity was assessed by comparing algorithm item scores with a clinical diagnosis made by two independent neuropsychiatrists or neuropsychologists familiar with ASDs. All of the subjects who received a clinical diagnosis of Asperger's Disorder or Atypical Autism ($n = 13$) met five or six of the algorithm criteria for Asperger's Disorder on at least one of the ratings. Of the remaining 11 individuals who were not diagnosed with Asperger's Disorder, only one met five criteria. The authors acknowledge that many of the individuals who met algorithm criteria for Asperger's Disorder would also meet *DSM-IV* criteria for autism. This instrument is in the preliminary stages

and further investigation is warranted prior to using it as a diagnostic instrument.

The Australian Scale for Asperger's Syndrome

The Australian Scale for Asperger's Syndrome (ASAS) was developed by Garnett and Attwood and published in Attwood's book, *Asperger's Syndrome: A guide for parents, professionals, people with Asperger's Syndrome and their partners* (Attwood, 1997), as well as on a web site (<http://www.tonyattwood.co.uk>). Although there are no peer reviewed published papers on this instrument, it is widely used by educational systems and parents, in large part because of the accessibility and popularity of the book and web site.

The ASAS covers five areas, which (as the developer of the instrument himself states) "loosely correspond to the five broad categories of behavior identified by other researchers to identify Asperger's Syndrome." These include social and emotional difficulties, cognitive skills deficits, communication skills deficits, specific interests, and motor clumsiness. The authors also indicate that there are at least two questions that are not based on current diagnostic criteria, because their clinical observations differed from what was reported in the literature. The instrument includes 19 items and is scored on a 7-point scale ranging from "rarely" (0) to "frequently" (6). Each item describes a behavior that the parent or teacher is asked to rate, followed by an example of that behavior.

A nonpeer reviewed study designed to evaluate the validity of the instrument in diagnosing Asperger's Disorder is available on Tony Attwood's web site. The study included children and adolescents between 3 and 19 years of age in three groups: a group of individuals referred to a clinic for Asperger's Disorder but not diagnosed with Asperger's Disorder, a group of individuals referred to a clinic and diagnosed with Asperger's Disorder, and a typical control group.

There are several concerns regarding the methods and design of this study. The ASAS was administered as an interview by a clinician, which is not the manner in which the instrument

is typically completed. It is intended to be completed as a questionnaire by a parent, teacher, or professional. Not only was it administered as an interview, but the interviewer was not blind to diagnosis. In addition, the clinical assessment consisted of "an unstructured clinical examination to decide whether they had AS." The assessment included a parent interview, an assessment with the child, a record review, and a diagnostic checklist. There were no standardized instruments used in the assessment. The relationship between the examiners in the Asperger's Clinic who made the diagnosis and the authors of the instrument was also unclear.

Based on a stepwise discriminant function analysis, accuracy for the predicted membership of the Asperger's Disorder group was 90%, and accuracy for the non-Asperger's group was 65%. Given its high sensitivity and low specificity, the authors recommend using this instrument as a screener, rather than as a diagnostic instrument at this time. They also caution against using the instrument clinically, given the lack of data on the reliability and validity of the instrument. Clearly, considering these results and the lack of carefully controlled studies, it is difficult to interpret results from the ASAS at this time.

Measuring Change in Core Behaviors

Investigators have often attempted to use diagnostic instruments in order to measure change in response to treatment. On the whole, this has not been very successful. This is partly due to the fact that most diagnostic instruments were designed to include a wide range of deficits associated with ASDs, and so they are not sufficiently sensitive to changes within an individual. In addition, expectations and contexts for behavior, especially for young children, frequently change with time (Lord et al., 2001; Volkmar & Lord, 1998). Although a child may be showing substantial improvement and acquiring specific behaviors, this improvement may not be measurable if the comparison is to the quality of interaction seen in typical children. On the other hand, for treatments that claim that they result in complete recovery, changes should be observable even in standard diagnostic instruments.

762 Assessment

There are a number of well-known instruments that measure behaviors that are not specific to autism but that are frequently found in association with it. These measures have often been used in psychopharmacology research. The most prominent one is the Aberrant Behavior Checklist (ABC; Aman, 1994; Aman & Singh, 1986; Arnold et al., 2000). The Autism Behavior Checklist (also known as the ABC; Krug et al., 1980a), although less appropriate as a diagnostic instrument, has also been helpful in indicating the degree of overtly abnormal or impairing behaviors produced, particularly by those children who are both autistic and mentally handicapped. The Children's Global Assessment Scale (Shaffer et al., 1983) gives a general measure of impairment, which may be helpful for some investigators. In addition, the Maladaptive Behavior Scale from the Vineland Adaptive Behavior Scale (Perry & Factor, 1989) provides counts of particular maladaptive behaviors. The Real Life Rating Scale (Freeman et al., 1986) has also been used for this aim. On the whole, most of these scales were not designed for diagnosis or measuring change and do not have psychometric data to support this particular use. The exception is the Aberrant Behavior Checklist. Recently, several investigators have begun to use the ADOS either as a measure or as a context in which to measure treatment responsiveness. In our own research, we see more quantifiable changes if we re-administer identical items over extended time periods (several years) on the direct observation schedules (e.g., ADOS), even given the variability that this entails, than we do in parent reports, because of the very broad focus of the ADI. Time will tell if the ADOS has a sufficient range of presses and contexts to be useful in this way (Owley et al., 2001).

CONCLUSIONS

Overall, there is a wealth of information and options for the diagnosis of autism, but there is still much to be done to make our techniques stronger and broader in scope. There will always be trade-offs between acquiring the maximum amount of meaningful information and highest validity versus being able to reliably code and make decisions about information. Users of diagnostic instruments

should be aware of the needs of their particular situation and population in order to make the most informed choice of instruments. In general, higher standards in terms of limiting the amount of information given to the user of an instrument tested (e.g., keeping examiners "blind" to diagnosis, attempting to use instruments with parents who have not yet received a diagnosis), and including measurements of test-retest reliability and appropriate analysis of reliability statistics, will aid in the interpretability of the instruments. Clear descriptions of exactly how instruments were used and are intended to be used, including cut-offs if categorical use is implied, are also critical.

It seems particularly important to recognize that there are a variety of needs having to do with formal diagnosis that may not be met by a single instrument. Screening of large populations for possible autism is most likely to occur with very young children and needs to be coordinated with developmental screening, because delays in language are inherently entwined with the recognition of autism in many children (see Stone, this volume). After a child has been identified as possibly having an ASD, procedures for early diagnosis may be rather different than screening methods. Diagnostic procedures will involve fewer children than screening and should have closer links to individual education and treatment plans, as well as outline possible multitaxial diagnoses.

For research purposes, there is a need for lifetime diagnoses and standard procedures that presumably yield the same final interpretation (though not necessarily the same raw data) for the same individual at multiple points in his or her life. In contrast, there is also a need for measurement of change. It seems very unlikely that any one instrument will accomplish all of these objectives. However, for each of these needs, there are promising candidates. Ensuring that the relationship between various instruments and goals is well understood will also increase the usefulness of the endeavor. Recognizing that other factors, particularly level of development and language skill, have marked effects on most measurements in autism and pervasive developmental disorder is an important step in considering the meaning of any clinical or research result.

Finally, there is a great need for the extension of the current instruments to diagnosis of disorders other than autism in the autism spectrum. Part of the difficulty, as discussed in later chapters, is that the definitions and discriminations from autism of these disorders are not yet as clear as we would like. However, reliable ways of formally substantiating diagnoses such as PDD-NOS, Asperger's Disorders and atypical autism are needed so that researchers and clinicians can make informed decisions about the usefulness of these concepts. Various instruments have been proposed to study these disorders, but at this point, they have little relationship to each other and have not been found to be reliable. Consequently, they offer limited scientific usefulness. A priority for researchers is to work together to derive operationalized definitions and specific proposals for how their approaches add to or fit in with those of other researchers. In the meantime, clinicians must be careful to be informed about the kind of information a particular instrument provides and to consider the implications for the appropriateness of that information to their immediate clinical needs.

Cross-References

Issues in the diagnosis of autism and related conditions are discussed in Chapters 1 to 7. Other aspects of assessment are reviewed in Chapter 27 and in Chapters 29 to 33.

REFERENCES

- Achenbach, T. M. (1981). *Childhood Behavior Checklist*. Burlington: University of Vermont, Department of Psychiatry.
- Achenbach, T. M., & Rescorla, L. (2000). *Manual for the ASEBA forms and profiles*. Burlington: University of Vermont, Center for Children, Youth and Families.
- Adrien, J. L., Perrot, A., Sauvage, D., Leddet, I., Larmande, C., Hameury, L., et al. (1992). Early symptoms in autism from family home movies: Evaluation and comparison between 1st and 2nd year of life using I.B.S.E. Scale. *Acta Paedopsychiatrica: International Journal of Child and Adolescent Psychiatry*, 55(2), 71–75.
- Alarcón, M., Cantor, R. M., Liu, J., Gilliam, T. C., Geschwind, D. H., & Autism Genetic Research Exchange Consortium. (2002). Evidence for a language quantitative trait locus on chromosome 7q in multiplex autism families. *American Journal of Human Genetics*, 70(1), 60–71.
- Aman, M. G. (1994). Instruments for assessing treatment effects in developmentally disabled populations. *Assessment in Rehabilitation and Exceptionality*, 1, 1–20.
- Aman, M. G., & Singh, N. N. (1986). *Aberrant Behavior Checklist: Manual*. East Aurora, NY: Slosson Educational Publications.
- American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.
- Arnold, L. E., Aman, M. G., Martin, A., Collier-Crespin, A., Vitiello, B., Tierney, E., et al. (2000). Assessment in multisite randomized clinical trials of patients with autistic disorder: The Autism RUPP Network. *Journal of Autism and Developmental Disorders*, 30(2), 99–111.
- Asano, E., Chugani, D. C., Muzik, O., Behen, M., Janisse, J., Rothermel, R., et al. (2001). Autism in tuberous sclerosis complex is related to both cortical and subcortical dysfunction. *Neurology*, 57(7), 1269–1277.
- Attwood, T. (1997). *Asperger's syndrome: A guide for parents and professionals*. London: Jessica Kingsley.
- Baron-Cohen, S., Allen, J., & Gillberg, C. (1992). Can autism be detected at 18 months? The needle, the haystack, and the CHAT. *British Journal of Psychiatry*, 161(1), 839–843.
- Baron-Cohen, S., Cox, A., Baird, G., Swettenham, J., Nightingale, N., Morgan, K., et al. (1996). Psychological markers in the detection of autism in infancy in a large population. *British Journal of Psychiatry*, 168(2), 158–163.
- Barthelemy, C., Adrien, J. L., Tanguay, P. E., Garreau, B., Fermanian, J., Roux, S., et al. (1990). The Behavioral Summarized Evaluation: Validity and reliability of a scale for the assessment of autistic behaviors. *Journal of Autism and Developmental Disorders*, 20(2), 189–204.
- Barthelemy, C., Roux, S., Adrien, J. L., Hameury, L., Guerin, P., Garreau, B., et al. (1997). Validation of the Revised Behavior Summarized Evaluation Scale. *Journal of Autism and Developmental Disorders*, 27(2), 139–153.
- Bebko, J. M., Konstantareas, M. M., & Springer, J. (1987). Parent and professional evaluations of family stress associated with characteristics of autism. *Journal of Autism and Developmental Disorders*, 17(4), 565–576.
- Berument, S. K., Rutter, M., Lord, C., Pickles, A., & Bailey, A. (1999). Autism screening questionnaire: Diagnostic validity. *British Journal of Psychiatry*, 175, 444–451.
- Bird, H. R., Gould, M. S., & Staghezza, B. (1992). Aggregating data from multiple informants in

- child psychiatry epidemiological research. *Journal of the American Academy of Child and Adolescent Psychiatry*, 31(1), 78–85.
- Bishop, D. V. M. (1998). Development of the Children's Communication Checklist (CCC): A method for assessing qualitative aspects of communicative impairment in children. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 39(6), 879–891.
- Bishop, D. V. M., & Baird, G. (2001). Parent and teacher report of pragmatic aspects of communication: Use of the Children's Communication Checklist in a clinical setting. *Developmental Medicine and Child Neurology*, 43(12), 809–818.
- 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 and Allied Disciplines*, 43(7), 917–929.
- Boiron, M., Barthelemy, C., Adrien, J. L., Martineau, J., & Lelord, G. (1992). The assessment of psychophysiological dysfunction in children using the BSE scale before and during therapy. *Acta Paedopsychiatrica: International Journal of Child and Adolescent Psychiatry*, 55(4), 203–206.
- Botting, N., & Conti-Ramsden, G. (1999). Pragmatic language impairment without autism: The children in question. *Autism*, 3(4), 371–396.
- Bradford, Y., Haines, J., Hutcheson, H., Gardiner, M., Braun, T., Sheffield, V., et al. (2001). Incorporating language phenotypes strengthens evidence of linkage to autism. *American Journal of Medical Genetics*, 105(6), 539–547.
- Buxbaum, J. D., Silverman, J. M., Smith, C. J., Kilifarski, M., Reichert, J., Hollander, E., et al. (2001). Evidence for a susceptibility gene for autism on chromosome 2 and for genetic heterogeneity. *American Journal of Human Genetics*, 68(6), 1514–1520.
- Chakrabarti, S., & Fombonne, E. (2001). Pervasive developmental disorders in preschool children. *Journal of the American Medical Association [Special issue]*, 285(24), 3093–3099.
- Charman, T., Swettenham, J., Baron-Cohen, S., Cox, A., Baird, G., & Drew, A. (1998). An experimental investigation of social-cognitive abilities in infants with autism: Clinical implications. *Infant Mental Health Journal*, 19(2), 260–275.
- Cicchetti, D. V., & Sparrow, S. S. (1981). Developing criteria for establishing inter-rater reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency*, 86(2), 127–137.
- Cohen, D. J., Caparulo, B. K., Gold, J. R., Waldo, M. C., Shaywitz, B. A., Rutterberg, B. A., et al. (1978). Agreement in diagnosis: Clinical assessment and behavior rating scales for pervasively disturbed children. *Journal of the American Academy of Child Psychiatry*, 17(4), 589–603.
- Constantino, J. N. (2002). *The Social Responsiveness Scale*. Los Angeles: Western Psychological Services.
- Constantino, J. N., Gruber, C. P., Davis, S., Hays, S., Passante, N., & Przybeck, T. (2004). The factor structure of autistic traits. *Journal of Child Psychology and Psychiatry*, 45(4), 719–726.
- Constantino, J. N., Przybeck, T., Friesen, D., & Todd, R. D. (2000). Reciprocal social behavior in children with and without pervasive developmental disorders. *Journal of Developmental and Behavioral Pediatrics*, 21(1), 2–11.
- Constantino, J. N., & Todd, R. D. (2000). Genetic structure of reciprocal social behavior. *American Journal of Psychiatry*, 157(12), 2043–2044.
- Constantino, J. N., & Todd, R. D. (2003). Autistic traits in the general population: A twin study. *Archives of General Psychiatry*, 60(5), 524–530.
- Cox, A., Klein, K., Charman, T., Baird, G., Baron-Cohen, S., Swettenham, J., et al. (1999). Autism spectrum disorders at 20 and 42 months of age: Stability of clinical and ADI-R diagnosis. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 40(5), 719–732.
- Cuccaro, M. L., Shao, Y. J., Bass, M. P., Abramson, R. K., Ravan, S. A., Wright, H. H., et al. (2003). Behavioral comparisons in autistic individuals from multiplex and singleton families. *Journal of Autism and Developmental Disorders*, 33(1), 87–91.
- Cuccaro, M. L., Shao, Y., Grubber, J., Slifer, M., Wolpert, C. M., Donnelly, S. L., et al. (2003). "Factor Analysis of Restricted and Repetitive Behaviors in Autism Using the Autism Diagnostic Interview-R." *Child Psychiatry and Human Development*, 34(1), 3–17.
- Dauids, A. (1975). Childhood psychosis: The problem of differential diagnosis. *Journal of Autism and Childhood Schizophrenia*, 5(2), 129–138.
- Dawson, G., Webb, S., Schellenberg, G. D., Dager, S., Friedman, S., Aylward, E., et al. (2002). Defining the broader phenotype of autism: Genetic, brain, and behavioral perspectives. *Development and Psychopathology*, 14(3), 581–611.
- de Bildt, A., Sytema, S., Ketelaars, C., Kraijer, D., Mulder, E., Volkmar, F., et al. (2004). Interrelationship between autism diagnostic observation schedule-generic (ADOS-G), autism diagnostic interview-revised (ADI-R), and the diagnostic and statistical manual of mental disorders (DSM-IV-TR) classification in chil-

- dren and adolescents with mental retardation. *Journal of Autism and Developmental Disorders* 34(2), 129–137.
- DiLavore, P., Lord, C., & Rutter, M. (1995). Pre-Linguistic Autism Diagnostic Observation Schedule (PLADOS). *Journal of Autism and Developmental Disorders*, 25(4), 355–379.
- Doll, E. A. (1965). *Vineland Social Maturity Scale*. Circle Pines, MN: American Guidance Service.
- Eaves, R. C. (1990). *The factor structure of autistic behavior*. Paper presented at the annual Alabama Conference on Autism, Birmingham.
- Eaves, R. C., Campbell, H. A., & Chambers, D. (2000). Criterion-related and construct validity of the Pervasive Developmental Disorders Rating Scale and the Autism Behavior Checklist. *Psychology in the Schools*, 37(4), 311–321.
- Eaves, R. C., & Hooper, J. (1987). A factor analysis of psychotic behavior. *Journal of Special Education*, 21(4), 122–132.
- Folstein, S. E., & Mankoski, R. E. (2000). Chromosome 7q: Where autism meets language disorder? *American Journal of Human Genetics*, 67(2), 278–281.
- Fombonne, E. (1992). Diagnostic assessment in a sample of autistic and developmentally impaired adolescents. *Journal of Autism and Developmental Disorders*, 22(4), 563–581.
- Freeman, N. L., Perry, A., & Factor, D. C. (1991). Child behavior as stressors: Replicating and extending the use of the CARS as a measure of stress: A research note. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 32(6), 1025–1030.
- Freeman, B. J., Ritvo, E. R., Guthrie, D., Schroth, P., & Ball, J. (1978). The Behavior Observation Scale for Autism: Initial methodology, data analysis, and preliminary findings on 89 children. *Journal of the American Academy of Child Psychiatry*, 17(4), 576–588.
- Freeman, B. J., Ritvo, E. R., Yokota, A., & Ritvo, A. (1986). A scale for rating symptoms of patients with the syndrome of autism in real life settings. *Journal of the American Academy of Child Psychiatry*, 25(1), 130–136.
- Freitag, C. M. (2002). Phenotypic characteristics of siblings with autism and/or pervasive developmental disorder: Evidence for heterogeneity. *American Journal of Medical Genetics*, 114(7), 31.
- Garfin, D. G., McCallon, D., & Cox, R. (1988). Validity and reliability of the Childhood Autism Rating Scale with autistic adolescents. *Journal of Autism and Developmental Disorders*, 18(3), 367–378.
- Ghaziuddin, M., Tsai, L., & Ghaziuddin, N. (1992). Brief report: A comparison of the diagnostic criteria for Asperger's Syndrome. *Journal of Autism and Developmental Disorders*, 22(4), 643–649.
- Gillberg, C., Gillberg, C., Rastam, M., & Wentz, E. (2001). The Asperger Syndrome (and high-functioning autism) Diagnostic Interview (ASDI): A preliminary study of a new structured clinical interview [Special issue]. *Autism*, 5(1), 57–66.
- Gilliam, J. E. (1995). *Gilliam Autism Rating Scale*. Austin, TX: ProEd.
- Grinager, A. N., Cox, N. J., & Yairi, E. (1997). The genetic basis of persistence and recovery in stuttering. *Journal of Speech and Hearing Research*, 40(3), 567–580.
- Happé, F. G. E. (1995). The role of age and verbal ability in the theory of mind task performance of subjects with autism. *Child Development*, 66(3), 843–855.
- Hepburn, S., John, A., Lord, C., & Rogers, S. (2003). *Sensitivity and specificity of the Autism Diagnostic Observation Schedule in young children*. Manuscript in preparation.
- Hertzog, M. E., Snow, M. E., New, E., & Shapiro, T. (1990). *DSM-III and DSM-III-R diagnosis of autism and pervasive developmental disorder in nursery school children*. *Journal of the American Academy of Child and Adolescent Psychiatry*, 29(1), 195–199.
- Hobson, R. P. (1991). Methodological issues for experiments on autistic individuals' perception and understanding of emotion. *Journal of the American Academy of Child and Adolescent Psychiatry*, 32(7), 1135–1158.
- Holroyd, S., & Baron-Cohen, S. (1993). Brief report: How far can people with autism go in developing a theory of mind? *Journal of Autism and Developmental Disorders*, 23(2), 379–385.
- Howlin, P., Mawhood, L., & Rutter, M. (2000). Autism and developmental receptive language disorder: A follow-up comparison in early adult life: Part II. Social, behavioural, and psychiatric outcomes. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 41(5), 561–578.
- International Molecular Genetic Study of Autism Consortium. (1998). A full genome screen for autism with evidence for linkage to a region on chromosome 7q. *Human Molecular Genetics*, 7(3), 571–578.
- Kanner, L. (1943). Autistic disturbances of affective contact. *Nervous Child*, 2, 217–250.
- Klin, A., Pauls, D. L., Schultz, R., & Volkmar, F. R. (in press). *Three diagnostic approaches to Asperger's syndrome: Implications for research*.
- Klin, A., Sparrow, S. S., Marans, W. D., Carter, A., & Volkmar, F. R. (2000). Assessment issues in

- children and adolescents with Asperger syndrome. In A. Klin, F. R. Volkmar, & S. S. Sparrow (Eds.), *Asperger syndrome* (pp. 309–339). New York: Guilford Press.
- Klin, A., Volkmar, F. R., Sparrow, S. S., Cicchetti, D. V., & Rourke, B. P. (1995). Validity and neuropsychological characterization of Asperger syndrome: Convergence with nonverbal learning disabilities syndrome. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 36(7), 1127–1140.
- Kolevzon, A., Smith, C. J., Schmeidler, J., Buxbaum, J. D., & Silverman, J. M. (2004). Familial symptom domains in monozygotic siblings with autism. *American Journal of Medical Genetics Part B-Neuropsychiatric Genetics* 129B, 76–81.
- Konstantareas, M. M., & Homatidis, S. (1989). Assessing child symptom severity and stress in parents of autistic children. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 30(3), 459–470.
- Kraemer, H. C. (1992). Measurement of reliability for categorical data medical research. *Statistical Methods in Medical Research*, 1(2), 183–199.
- Krug, D. A., Arick, J. R., & Almond, P. J. (1980a). *Autism screening instrument for educational planning*. Portland, OR: ASIEP Educational.
- Krug, D. A., Arick, J. R., & Almond, P. J. (1980b). Behavior checklist for identifying severely handicapped individuals with high levels of autistic behavior. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 21(3), 221–229.
- Krug, D. A., Arick, J. R., & Almond, P. J. (1993). *Autism screening instrument for educational planning* (2nd ed.). Austin, TX: ProEd.
- Kurita, H., Kita, M., & Miyake, Y. (1992). A comparative study of development and symptoms among disintegrative psychosis and infantile autism with and without speech loss. *Journal of Autism and Developmental Disorders*, 22(2), 175–188.
- Lam, M. K., & Rao, N. (1993). Developing a Chinese version of the Psychoeducational Profile (CPEP) to assess autistic children in Hong Kong. *Journal of Autism and Developmental Disorders*, 23(2), 273–279.
- Le Couteur, A., Bailey, A., Goode, S., Pickles, A., Robertson, S., Gottesman, I., et al. (1996). A broader phenotype of autism: The clinical spectrum in twins. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 37(7), 785–801.
- Le Couteur, A., Lord, C., & Rutter, M. (2003). *The Autism Diagnostic Interview: Revised* (ADI-R). Los Angeles: Western Psychological Services.
- Le Couteur, A., Rutter, M., Lord, C., Rios, P., Robertson, S., Holdgrafer, M., et al. (1989). Autism Diagnostic Interview: A standardized investigator-based instrument. *Journal of Autism and Developmental Disorders*, 19(3), 363–387.
- Leekam, S. R., Libby, S. J., Wing, L., Gould, J., & Gillberg, C. (2000). Comparison of ICD-10 and Gillberg's criteria for Asperger syndrome [Special issue: Asperger syndrome]. *Autism*, 4(1), 11–28.
- Leekam, S. R., Libby, S. J., Wing, L., Gould, J., & Taylor, C. (2002). The Diagnostic Interview for Social and Communication Disorders: Algorithms for ICD-10 childhood autism and Wing and Gould autistic spectrum disorder. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 43(3), 327–342.
- Lord, C. (1990). A cognitive-behavioral model for the treatment of social-communicative deficits in adolescents with autism. In R. J. McMahon & R. D. Peters (Eds.), *Behavior disorders of adolescence: Research, intervention and policy in clinical and school settings* (pp. 155–174). New York: Plenum Press.
- Lord, C. (1995). Follow-up of two-year-olds referred for possible autism. *Journal of Child Psychology and Psychiatry*, 36(8), 1365–1382.
- Lord, C. (1996). Treatment of a high-functioning adolescent with autism: A cognitive-behavioral approach. *Cognitive therapy with children and adolescents: A casebook for clinical practice* (pp. 394–404). New York: Guilford Press.
- Lord, C., & Bailey, A. (2002). Autism spectrum disorders. In M. Rutter & E. Taylor (Eds.), *Child and adolescent psychiatry* (4th ed., pp. 636–663). Oxford, England: Blackwell.
- Lord, C., Cook, E. H., Leventhal, B. L., & Amaral, D. G. (2000). Autism spectrum disorders. *Neuron*, 28(2), 355–363.
- Lord, C., Leventhal, B. L., & Cook, E. H., Jr. (2001). Quantifying the phenotype in autism spectrum disorders. *American Journal of Medical Genetics*, 105(1), 36–38.
- Lord, C., Pickles, A., DiLavore, P. C., & Shulman, C. (1996). *Longitudinal studies of young children referred for possible autism*. Paper presented at the biannual meeting of the International Society for Research in Child and Adolescent Psychopathology, Los Angeles.
- Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Jr., Leventhal, B. L., DiLavore, P. C., et al. (2000). The Autism Diagnostic Observation Schedule-Generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders*, 30(3), 205–223.

- Lord, C., Rutter, M. L., DiLavore, P. C., & Risi, S. (1999). *Autism Diagnostic Observation Schedule—WPS* (WPS ed.). Los Angeles: Western Psychological Services.
- Lord, C., Rutter, M. L., Goode, S., Heemsbergen, J., Jordan, H., Mawhood, L., et al. (1989). Autism Diagnostic Observation Schedule: A standardized observation of communicative and social behavior. *Journal of Autism and Developmental Disorders*, *19*(2), 185–212.
- Lord, C., Rutter, M. L., & Le Couteur, A. (1994). The Autism Diagnostic Interview—Revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders*, *24*(5), 659–685.
- Lord, C., Storoschuk, S., Rutter, M., & Pickles, A. (1993). Using the ADI-R to diagnose autism in preschool children. *Infant Mental Health Journal*, *14*(3), 1234–1252.
- Luteijn, E., Luteijn, F., Jackson, S., Volkmar, F., & Minderaa, R. (2000). The Children's Social Behavior Questionnaire for milder variants of PDD problems: Evaluation of the psychometric characteristics. *Journal of Autism and Developmental Disorders*, *30*(4), 317–330.
- MacLean, J. E., Szatmari, P., Jones, M. B., Bryson, S. E., Mahoney, W. J., Bartolucci, G., et al. (1999). Familial factors influence level of functioning in pervasive developmental disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, *38*(6), 746–753.
- Mahoney, W. J., Szatmari, P., MacLean, J. E., Bryson, S. E., Bartolucci, G., Walter, S. D., et al. (1998). Reliability and accuracy of differentiating pervasive developmental disorder subtypes. *Journal of the American Academy of Child and Adolescent Psychiatry*, *37*(3), 278–285.
- Masters, J. C., & Miller, D. E. (1970). Early infantile autism: A methodological critique. *Journal of Abnormal Psychology*, *75*(3), 342–343.
- Mawhood, L., Howlin, P., & Rutter, M. L. (2000). Autism and developmental receptive language disorder—A comparative follow-up in early adult life: Part I. Cognitive and language outcomes. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *41*(5), 547–559.
- Mayes, L. C., & Zigler, E. (1992). An observational study of the affective concomitants of mastery in infants. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *33*(4), 659–667.
- Mesibov, G. B., Schopler, E., & Caison, W. (1989). The Adolescent and Adult Psychoeducational Profile: Assessment of adolescents and adults with severe developmental handicaps. *Journal of Autism and Developmental Disorders*, *19*(1), 33–40.
- Mesibov, G. B., Schopler, E., Schaffer, B., & Landrus, R. (1988). *Adolescent and Adult Psychoeducational Profile* (AAPEP): Volume IV. Austin, TX: ProEd.
- Mesibov, G. B., Schopler, E., Schaffer, B., & Michal, N. (1989). Use of the Childhood Autism Rating Scale with autistic adolescents and adults. *Journal of the American Academy of Child and Adolescent Psychiatry*, *28*(4), 538–541.
- Miller, J. N., & Ozonoff, S. (1997). Did Asperger's cases have Asperger disorder? *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *38*(2), 247–251.
- Minshew, N. J., & Goldstein, G. (1993). Is autism an amnesic disorder? Evidence from the California Verbal Learning Test. *Neuropsychology*, *7*(2), 209–216.
- Miranda-Linne, F. M., Fredrika, M., & Melin, L. (1997). A comparison of speaking and mute individuals with autism and autistic-like conditions on the Autism Behavior Checklist. *Journal of Autism and Developmental Disorders*, *27*(3), 245–264.
- Mundy, P., Sigman, M., Ungerer, J., & Sherman, T. (1986). Defining the social deficits of autism: The contribution of nonverbal communication measures. *Journal of Child Psychology and Psychiatry*, *27*(5), 657–669.
- Munson, J., Dawson, G., Lord, C., Rogers, S., & Sigman, M. (in press). *Cognitive profiles and adaptive functioning in preschool children with autism spectrum disorder versus developmental delay*.
- Muris, P., Steerneman, P., & Ratering, E. (1997). Inter-rater reliability of the Psychoeducational Profile (PEP). *Journal of Autism and Developmental Disorders*, *27*(5), 621–626.
- Nordin, V., & Gillberg, C. (1996a). Autism spectrum disorders in children with physical or mental disability or both: Part I. Clinical and epidemiological aspects. *Developmental Medicine and Child Neurology*, *38*(4), 297–313.
- Nordin, V., & Gillberg, C. (1996b). Autism spectrum disorders in children with physical or mental disability or both: Part II. Screening aspects. *Journal of Child Psychology and Psychiatry*, *38*(4), 314–324.
- Nordin, V., & Gillberg, C. (1998). The long-term course of autistic disorders: Update on follow-up studies. *Acta Psychiatrica Scandinavica*, *97*(2), 99–108.
- Nordin, V., Gillberg, C., & Nyden, A. (1998). The Swedish version of the Childhood Autism

- Rating Scale in a clinical setting. *Journal of Autism and Developmental Disorders*, 28(1), 69–75.
- Offord, D. R., Boyle, M. H., Racine, Y., Szatmari, P., Fleming, J. E., Sanford, M., et al. (1996). Integrating assessment data from multiple informants. *Journal of the American Academy of Child and Adolescent Psychiatry*, 35(8), 1078–1085.
- Owley, T., McMahon, W., Cook, E. H., Laulhere, T., South, M., Mays, L. Z., et al. (2001). Multi-site, double-blind, placebo-controlled trial of porcine secretin in autism. *Journal of the American Academy of Child and Adolescent Psychiatry*, 40(11), 1293–1299.
- Ozonoff, S., & Cathcart, K. (1998). Effectiveness of a home program intervention for young children with autism. *Journal of Autism and Developmental Disorders*, 28(1), 25–32.
- Ozonoff, S., South, M., & Miller, J. N. (2000). DSM-IV-defined Asperger syndrome: Cognitive, behavioral and early history differentiation from high-functioning autism [Special issue: Asperger syndrome]. *Autism*, 4(1), 29–46.
- Panerai, S., Ferrante, L., & Caputo, V. (1997). The TEACCH strategy in mentally retarded children with autism: A multidimensional assessment: Pilot study. *Journal of Autism and Developmental Disorders*, 27(3), 345–347.
- Parks, S. L. (1988). Psychometric instruments available for the assessment of autistic children. In E. Schopler & G. Mesibov (Eds.), *Diagnosis and assessment in autism* (pp. 123–136). New York: Plenum Press.
- Perry, A., & Factor, D. C. (1989). Psychometric validity and clinical usefulness of the Vineland Adaptive Behavior Scale and the AAMD Adaptive Behavior Scale for an autistic sample. *Journal of Autism and Developmental Disorders*, 19(1), 41–56.
- Persson, B. (2000). Brief report: A longitudinal study of quality of life and independence among adult men with autism. *Journal of Autism and Developmental Disorders*, 30(1), 2061–2066.
- Pilowsky, T., Yirmiya, N., Shulman, C., & Dover, R. (1998). The Autism Diagnostic Review–Revised and the childhood autism rating scale: Differences between diagnostic systems and comparison between genders. *Journal of Autism and Developmental Disorders*, 28(2), 143–151.
- Piven, J., Harper, J., Palmer, P., & Arndt, S. (1996). Course of behavioral change in autism: A retrospective study of high-IQ adolescents and adults. *Journal of the American Academy of Child and Adolescent Psychiatry*, 35(4), 523–529.
- Piven, J., Nehme, E., Simon, J., Barta, P., Pearl, G., & Folstein, S. E. (1992). Magnetic Resonance Imaging in autism: Measurement of the cerebellum, pons, and fourth ventricle. *Biological Psychiatry*, 31(5), 491–504.
- Prior, M. R., & Bence, R. (1975). A note on the validity of the Rimland Diagnostic Checklist. *Journal of Clinical Psychology*, 31(3), 510–513.
- Rimland, B. (1968). On the objective diagnosis of infantile autism. *Acta Paedopsychiatrica: International Journal of Child and Adolescent Psychiatry*, 35(4/8), 146–161.
- Rimland, B. (1971). The differentiation of childhood psychoses: An analysis of checklists for 2,218 psychotic children. *Journal of Autism and Childhood Schizophrenia*, 1(2), 161–174.
- Robertson, J. M., Tanguay, P. E., L'Ecuyer, S., Sims, A., & Waltrip, C. (1999). Domains of social communication handicap in autism spectrum disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, 38(6), 738–745.
- Ruttenberg, B. A., Dratman, M. L., Fraknoi, J., & Wenar, C. (1966). An instrument for evaluating autistic children. *Journal of American Academy of Child Psychiatry*, 5, 453–478.
- Ruttenberg, B. A., Kalish, B. I., Wenar, C., & Wolf, E. G. (1977). *Behavior rating instrument for autistic and other atypical children* (rev. ed.). Philadelphia: Developmental Center for Autistic Children.
- Rutter, M., Le Couteur, A., & Lord, C. (2003). *Manual for the ADI–WPS version*. Los Angeles: Western Psychological Services.
- Rutter, M., Mawhood, L., & Howlin, P. (1992). Language delay and social development. In P. Fletcher & D. Hall (Eds.), *Specific speech and language disorders in children: Correlates, characteristics, and outcomes* (pp. 63–78). London: Whurr.
- Saemundsen, E., Magnússon, P., Smári, J., & Sigurdardóttir, S. (2003). Autism Diagnostic Interview–Revised and the Childhood Autism Rating Scale: Convergence and discrepancy in diagnosing autism. *Journal of Autism and Developmental Disorders* 33(3), 319–328.
- Sanchez, L. E., Adams, P. B., Yusal, S., Hallin, A., Campbell, M., & Small, A. M. (1995). A comparison of live and videotape ratings: Comipramine and halperidol in autism. *Psychopharmacology Bulletin*, 31(2), 371–378.
- Schopler, E. (1976). Towards reducing behavior problems in autistic children. In L. Wing (Ed.), *Early childhood autism* (pp. 221–246). London: Pergamon Press.
- Schopler, E., & Reichler, R. J. (1972). How well do parents understand their own psychotic child?

- Journal of Autism and Childhood Schizophrenia*, 2(4), 387–400.
- Schopler, E., & Reichler, R. J. (1979). *Individualized assessment and treatment for autistic and developmentally disabled children: Psychoeducational profile* (Vol. 1). Baltimore: University Park Press.
- Schopler, E., Reichler, R. J., Bashford, A., Lansing, M. D., & Marcus, L. M. (1990). *Psychoeducational Profile-Revised*. Austin, TX: ProEd.
- Schopler, E., Reichler, R. J., DeVellis, R., & Daly, K. (1980). Toward objective classification of childhood autism: Childhood Autism Rating Scale (CARS). *Journal of Autism and Developmental Disorders*, 10(1), 91–103.
- Schopler, E., Reichler, R. J., & Renner, B. R. (1986). *The Childhood Autism Rating Scale (CARS) for diagnostic screening and classification of autism*. Irvington, NY: Irvington.
- Schopler, E., Reichler, R. J., & Renner, B. R. (1988). *The Childhood Autism Rating Scale (CARS)*. Los Angeles: Western Psychological Services.
- Schreck, K. A., & Mulick, J. A. (2000). Parental report of sleep problems in children with autism. *Journal of Autism and Developmental Disorders*, 30(2), 127–135.
- Sevin, J. A., Matson, J. L., Coe, D. A., Fee, V. E., & Sevin, B. M. (1991). A comparison and evaluation of three commonly used autism scales. *Journal of Autism and Developmental Disorders*, 21(4), 551–556.
- Shaffer, D., Gould, M. S., Brasic, J., Ambrosini, P., Fisher, P., Bird, H., et al. (1983). A Children's Global Assessment Scale (CGAS). *Archives of General Psychiatry*, 40, 1228–1231.
- Shao, Y. J., Cuccaro, M. L., Hauser, E. R., Raiford, K. L., Menold, M. M., Wolpert, C. M., et al. (2003). Fine mapping of Autistic disorder to chromosome 15q11-q13 by use of phenotypic subtypes. *American Journal of Human Genetics*, 72(3), 539–548.
- Shao, Y. J., Raiford, K. L., Wolpert, C. M., Cope, H. A., Ravan, S. A., Ashley-Koch, A. A., et al. (2002). Phenotypic homogeneity provides increased support for linkage on chromosome 2 in autistic disorder. *American Journal of Human Genetics*, 70(4), 1058–1061.
- Shao, Y. J., Wolpert, C. M., Raiford, K. L., Menold, M. M., Donnelly, S. L., Ravan, S. A., et al. (2002). Genomic screen and follow-up analysis for autistic disorder. *American Journal of Medical Genetics*, 114, 99–105.
- Sigman, M., & Ungerer, J. (1984). Attachment behaviors in autistic children. *Journal of Autism and Developmental Disorders*, 14(3), 231–244.
- Silverman, J. M., Smith, C. J., Schmeidler, J., Hollander, E., Lawlor, B. A., Fitzgerald, M., et al. (2002). Symptom domains in autism and related conditions: Evidence for familiarity. *American Journal of Medical Genetics*, 114(1), 64–73.
- Smalley, S. L., Tanguay, P. E., Smith, M., & Gutierrez, G. (1992). Autism and tuberous sclerosis. *Journal of Autism and Developmental Disorders*, 22(3), 339–355.
- South, M., Williams, B. J., McMahon, W. M., Owley, T., Filipek, P. A., Shernoff, E., et al. (2002). Utility of the Gilliam Autism Rating Scale in research and clinical populations. *Journal of Autism and Developmental Disorders*, 32(6), 593–599.
- Sparrow, S. S., Balla, D., & Cicchetti, D. (1984). *Vineland Adaptive Behavior Scales*. Circle Pines, MN: American Guidance Service.
- Spencer, A. (1993). *Separation and reunion in autistic two year olds*. Unpublished doctoral dissertation, University of North Carolina, Chapel Hill.
- Spiker, D., Lotspeich, L. J., Dimiceli, S., Myers, R. M., & Risch, N. (2002). Behavioral phenotypic variation in autism multiplex families: Evidence for a continuous severity gradient. *American Journal of Medical Genetics*, 114(2), 129–136.
- Spiker, D., Lotspeich, L. J., Kraemer, H. C., Hallmayer, J., McMahon, W., Peterson, B., et al. (1994). Genetics of autism: Characteristics of affected and unaffected children from 37 multiplex families. *American Journal of Medical Genetics*, 54(1), 27–35.
- Sponheim, E. (1996). Changing criteria of autistic disorders: A comparison of the ICD-10 research criteria and DSM-IV with DSM-III-R, CARS, and ABC. *Journal of Autism and Developmental Disorders*, 26(5), 513–525.
- Steerneman, P., Muris, P., Merckelbach, H., & Willems, H. (1997). Brief report: Assessment of development and abnormal behavior in children with pervasive developmental disorders: Evidence for the reliability and validity of the Revised Psychoeducational Profile. *Journal of Autism and Developmental Disorders*, 27(2), 177–185.
- Stella, J., Mundy, P., & Tuchman, R. (1999). Social and nonsocial factors in the Childhood Autism Rating Scale. *Journal of Autism and Developmental Disorders*, 29(4), 307–317.
- Stone, W. L., Lee, E. B., Ashford, L., Brissie, J., Hepburn, S. L., Coonrod, E. E., et al. (1999). Can autism be diagnosed accurately in children under 3 years? *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 40(2), 219–226.
- Stone, W. L., & Lemanek, K. L. (1990). Parental report of social behaviors in autistic

- preschoolers. *Journal of Autism and Developmental Disorders*, 20(4), 513–522.
- Stone, W. L., Ousley, O. Y., Yoder, P., Hogan, K., & Hepburn, S. (1997). Nonverbal communication in 2- and 3-year-old children with autism. *Journal of Autism and Developmental Disorders*, 27(6), 677–696.
- Sturme, P., Matson, J. L., & Sevin, J. A. (1992). Brief report: Analysis of the internal consistency of three autism scales. *Journal of Autism and Developmental Disorders*, 22(2), 321–328.
- Szatmari, P. (2000). Perspectives on the classification of Asperger Syndrome. In A. Klin (Ed.), *Asperger Syndrome* (pp. 403–407). New York, NY: Guilford Press.
- Szatmari, P., Archer, L., Fisman, S., Streiner, D. L., & Wilson, F. (1995). Asperger's syndrome and autism: Differences in behavior, cognition, and adaptive functioning. *Journal of the American Academy of Child and Adolescent Psychiatry*, 34(12), 1662–1671.
- Szatmari, P., Merette, C., Bryson, S. E., Thivierge, J., Roy, M. A., Cayer, M., et al. (2002). Quantifying dimensions in autism: A factor-analytic study. *Journal of the American Academy of Child and Adolescent Psychiatry*, 41(4), 467–474.
- Tadevosyan-Leyfer, O., Dowd, M., Mankoski, R., Winklosky, B., Putnam, S., McGrath, L., et al. (2003). A principal components analysis of the Autism Diagnostic Interview-Revised. *Journal of the American Academy of Child and Adolescent Psychiatry*, 42(7), 864–872.
- Tanguay, P. E., Robertson, J., & Derrick, A. (1998). A dimensional classification of autism spectrum disorder by social communication domains. *Journal of the American Academy of Child and Adolescent Psychiatry*, 37(3), 271–277.
- Tantam, D. (2000). Psychological disorder in adolescents and adults with Asperger Syndrome. *Autism*, 4(1), 47–62.
- Teal, M. B., & Wiebe, M. J. (1986). A validity analysis of selected instruments used to assess autism. *Journal of Autism and Developmental Disorders*, 16(4), 485–494.
- Van Bourgondien, M. E., Marcus, L. M., & Schopler, E. (1992). Comparison of *DSM-III-R* and Childhood Autism Rating Scale diagnosis of autism. *Journal of Autism and Developmental Disorders*, 22(4), 493–505.
- Venter, A., Lord, C., & Schopler, E. (1992). A follow-up study of high-functioning autistic children. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 33(3), 1489–1507.
- Volkmar, F. R., Cicchetti, D. V., Bregman, J., & Cohen, D. J. (1992). Three diagnostic systems for autism: *DSM-III*, *DSM-III-R*, and *ICD-10* [Special issue: Classification and diagnosis]. *Journal of Autism and Developmental Disorders*, 22(4), 483–492.
- Volkmar, F. R., Cicchetti, D. V., Dykens, E., Sparrow, S. S., Leckman, J. F., & Cohen, D. F. (1988). An evaluation of the Autism Behavior Checklist. *Journal of Autism and Developmental Disorders*, 18(1), 81–97.
- Volkmar, F. R., & Klin, A. (2001). Asperger's disorder and higher functioning autism: Same or different? In L. M. Glidden (Ed.), *International review of research in mental retardation: Autism* (Vol. 23, pp. 83–110). San Diego, CA: Academic Press.
- Volkmar, F. R., Klin, A., Siegal, B., Szatmari, P., Lord, C., Campbell, M., et al. (1994). Field trial for autistic disorder in *DSM-IV*. *American Journal of Psychiatry*, 151(9), 1361–1367.
- Volkmar, F. R., & Lord, C. (1998). Diagnosis and definition of autism and other pervasive developmental disorders. In F. R. Volkmar (Ed.), *Autism and pervasive developmental disorders* (pp. 1–31). New York: Cambridge University Press.
- Vrancic, D., Nanclares, V., Soares, D., Kulesz, A., Mordzinski, C., Plebst, C., et al. (2002). Sensitivity and specificity of the autism diagnostic inventory-telephone screening in Spanish. *Journal of Autism and Developmental Disorders*, 32(4), 313–320.
- Wadden, N., Bryson, S. E., & Rodger, R. (1991). A closer look at the Autism Behavior Checklist: Discriminant validity and factor structure. *Journal of Autism and Developmental Disorders*, 21(4), 529–542.
- Wechsler, D. (1991). *Manual for the Wechsler Intelligence Scale for Children-III*. San Antonio, TX: Psychological Corporation.
- Wechsler, D. (1997). *Wechsler Adult Intelligence Scale, 3rd Edition*. San Antonio, TX: Psychological Corporation.
- Wechsler, D. (2002). *Wechsler Preschool and Primary Scale of Intelligence-III*. San Antonio, TX: Psychological Corporation.
- Wechsler, D. (2003). *Wechsler Intelligence Scale for Children (4th ed.)*. San Antonio, TX: Psychological Corporation.
- Wenar, C., & Rutenber, B. A. (1976). The use of BRIAAC for evaluating therapeutic effectiveness. *Journal of Autism and Childhood Schizophrenia*, 6(2), 175–191.
- Wetherby, A., & Prizant, B. (1993). *Communication and Symbolic Behavior Scales* (Normed ed.). Baltimore: Paul H. Brookes.

- Wetherby, A., & Prizant, B. (2002). *Communication and Symbolic Behavior Scales developmental profile* (First Normed Edition). Baltimore: Paul H. Brookes.
- Wetherby, A., Woods, J., Allen, L., Cleary, J., Dickinson, H., & Lord, C. (in press). *Early indicators of autistic spectrum disorders in the second year of life*.
- Wing, L., & Attwood, A. (1987). Syndromes of autism and atypical development. In D. J. Cohen & A. M. Donnellan (Eds.), *Handbook of autism and pervasive developmental disorders* (pp. 148–170). New York: Wiley.
- Wing, L., & Gould, J. (1978). Systematic recording of behaviors and skills of retarded and psychotic children. *Journal of Autism and Childhood Schizophrenia*, 8(1), 79–97.
- Wing, L., & Gould, J. (1979). Severe impairments of social interaction and associated abnormalities in children: Epidemiology and classification. *Journal of Autism and Developmental Disorders*, 9(1), 11–29.
- Wing, L., Leekam, S. R., Libby, S. J., Gould, J., & Locombe, M. (2002). The Diagnostic Interview for Social and Communication Disorders: Background, inter-rater reliability and clinical use. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 43(3), 307–325.
- World Health Organization. (1992). *The ICD 10 Classification of Mental and Behavioral Disorders: Clinical descriptions and diagnostic guidelines*. Geneva, Switzerland: Author.
- Yirmiya, N., Sigman, M., & Freeman, B. J. (1994). Comparison between diagnostic instruments for identifying high-functioning children with autism. *Journal of Autism and Developmental Disorders*, 24(3), 281–291.
- Zakian, A., Malvy, J., Desombre, H., Roux, S., & Lenoir, P. (2000). Early signs of autism: A new study of family home movies. *Encephale-Revue De Psychiatrie Clinique Biologique Et Therapeutique*, 26(2), 38–44.