Journal of Autism and Developmental Disorders, Vol. 35, No. 1, February 2005 (© 2005)

Diagnostic Assessment of Asperger's Disorder: A Review of Five Third-Party Rating Scales

Jonathan M. Campbell[1]

Five rating scales for screening and detection of Asperger's Disorder, three commercially available and two research instruments, are evaluated with reference to psychometric criteria outlined by Bracken in 1987 (*Journal of Psychoeducational Assessment*, 4, 313). Reliability and validity data reported in examiner's manuals or published reports are reviewed. The scales included in the review are the Asperger Syndrome Diagnostic Scale (ASDS), Autism Spectrum Screening Questionnaire (ASSQ), Childhood Asperger Syndrome Test (CAST), Gilliam Asperger's Disorder Scale (GADS), and Krug Asperger's Disorder Index (KADI). All published rating scales demonstrated significant weaknesses, particularly in the use of questionable normative samples. Among the published instruments, the KADI appears to be the most sound in terms of reliability and validity. The research instruments present incomplete psychometric data to date, but hold promise as clinical instruments.

KEY WORDS: Asperger's Disorder; assessment; validity; rating scales.

Asperger's Disorder (AD) is a developmental disorder characterized by significant impairments in social communication and restricted patterns of interest or behaviors in the presence of generally age-appropriate language acquisition and cognitive functioning (Klin, Volkmar, & Sparrow, 2000). AD was first described in a series of case studies by Hans Asperger (Asperger, 1944) and "reintroduced" with Lorna Wing's (1981) description of a series of clinical cases. Interest in AD grew after Wing's clinical account and has culminated in the inclusion of AD into widely used diagnostic classification systems, such as the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV; American Psychiatric Association, 1994), its revision (DSM-IV-TR; American Psychiatric Association, 2000), and the International Classification of Diseases and Related Health Problems, Tenth Edition (ICD-10; World Health Organization, 1992). Despite ongoing controversy regarding the validity of AD as separate from autism, particularly "high functioning" autism (e.g., Campbell & Morgan, 1998; Mayes, Calhoun, & Crites, 2001), a number of measures have been developed to detect and screen for AD (Howlin, 2000).

The purpose of the present review is to evaluate third-party rating scales currently available for use in assisting in the detection and evaluation of individuals with AD. The author describes each rating scale including its format, length, and intended purposes. The description is followed by a review of each scale's construction and psychometric properties, such as various forms of reliability and validity as described below. Each scale is then evaluated in terms of how well it fulfills psychometric criteria established by Bracken (1987), which appears in the form of a short critique for each measure.

[1] Jonathan M. Campbell, Ph.D., Assistant Professor, Department of Educational Psychology, The University of Georgia, 329 Aderhold Hall, Athens, GA 30602-7143; e-mail: jcampbel@coe.uga.edu

METHOD

Rating Scales

The five rating scales selected for review consist of three commercially available measures, the Asperger Syndrome Diagnostic Scale (ASDS; Myles, Bock, & Simpson, 2001), Gilliam Asperger's Disorder Scale (GADS; Gilliam, 2001), and Krug Asperger's Disorder Index (KADI; Krug & Arick, 2003), and two research instruments, the Autism Spectrum Screening Questionnaire (ASSQ; Ehlers, Gillberg, & Wing, 1999) and Childhood Asperger's Screening Test (CAST; Scott, Baron-Cohen, Bolton, & Brayne, 2002). The author conducted literature searches and reviewed recently published catalogues in selecting commercially available measures of AD. Literature reviews were also conducted to locate additional third-party rating scales designed to detect, screen, or identify individuals with AD. Measures reviewed but not included in the present paper were the Autism Screening Questionnaire (ASQ; Berument, Rutter, Lord, Pickles, & Bailey, 1999), the Ghuman–Folstein Screen for Social Interaction (SSI; Ghuman, Freund, Reiss, Serwint, & Folstein, 1998), and the Australian Scale for Asperger Syndrome (ASAS; Attwood, 1998). The ASQ and SSI were not included in the review as the measures were not designed to screen specifically for AD; the ASAS was not included in the review as no research reports have been published to document psychometric properties of the scale.

Procedure

Examiner's manuals from each of the published tests and published reports for the research instruments were evaluated according to Bracken's (1987) standards of psychometric adequacy. Bracken outlined a useful set of psychometric criteria that have been used to evaluate the technical adequacy of measures assessing cognitive, language, and social-emotional functioning (e.g., Bracken, Keith, & Walker, 1994; Campbell, 1998). Bracken's criteria state that technically adequate measures should be internally consistent and stable over time as evidenced by: (a) an internal consistency value of .90 or greater for a test's total score, (b) a median subtest internal consistency value of .80 or greater, and (c) a total test temporal stability value of .90 or greater.

Bracken also stated that technically adequate measures should show an appropriate range of standard and scaled scores (e.g., adequate floors and ceilings) in order to differentiate between typical and atypical development. For the purposes of behavior rating scales, Bracken *et al.* (1994) defined adequate test floors and ceilings as: (a) total test score floors and ceilings at least two standard deviations beyond the normative mean, and (b) average subtest scaled score floors and ceilings at least two standard deviations beyond the normative mean. The Bracken guidelines further outlined the importance of item gradients for instruments, that is, the degree to which standard or scaled scores change as a function of a child's performance on a single item. The larger the change in standard (or scaled) score as a function of a single item, the less sensitive the measure is to detect small differences in the child's functioning. Per Bracken, total test and average subtest item gradients should be no steeper than 1/3 of a standard deviation (i.e., an item should count no more than 1/3 of a standard deviation for a total standard score or an average subtest scaled score). Finally, Bracken defines technically adequate instrumentation by the presence of validity data.

In the case of rating scales, Streiner (1993) further recommended that rating items be comprehensible by lay respondents. Therefore, instructions and items for each rating scale were typed into Microsoft Word and readability statistics, i.e., Flesh Reading Ease (FRE) and Flesh-Kincaid Grade Level (F-KG), were calculated. The FRE rates text on a 100-point scale with higher scores indicating easier comprehension; the F-KG score rates text on a grade-based level (e.g., F-KG score of 7.0 indicates that text is comprehensible for the average seventh grader). Results of the review for each test are presented in alphabetical order.

RESULTS

Asperger Syndrome Diagnostic Scale (ASDS)

Brief Description of Test

The ASDS is a 50-item norm-referenced rating scale that requires the respondent to indicate the presence or absence of behaviors indicative of AD. The ASDS contains five subscales: language, social, maladaptive, cognitive, and sensorimotor (see Table I). Raw scores are summed within the five domains and yield scaled scores ($M = 10$, $SD = 3$) and percentile ranks for subtests. Items are summed for the entire scale to yield an Asperger Syndrome Quotient (ASQ), which is a standard score ($M = 100$,

Table I. Description of Third-Party Ratings for Asperger's Disorder

	Purpose(s)	Ages[a]	Items	Subscales	Readability (RE/GL)[b]
ASDS	Diagnostic aid Monitor behavior Generate IEP goals Research	5–18	50	Language Social Maladaptive Cognitive Sensorimotor	40.5/9.9
ASSQ	Screener Research	6–17	27	Overall score	46.3/8.7
CAST	Screener Research	4–11	37	Overall score	74.9/4.1
GADS	Diagnostic aid Assess unique behavior Monitor behavior Generate IEP goals Research	3–22	32	Social interaction Restricted behaviors Cognitive patterns Pragmatic skills	55.7/8.1
KADI	Screener Generate IEP goals Research	6–21	32	Overall score	51.2/8.2

Note: ASDS – Asperger Syndrome Diagnostic Scale; ASSQ – Autism Spectrum Screening Questionnaire; CAST – Childhood Asperger Syndrome Test; GADS – Gilliam Asperger's Disorder Scale; KADI – Krug Asperger's Disorder Index.

[a] Age ranges presented in years.

[b] RE – Reading ease represented by the Flesch Reading Ease score which rates text on a 100-point scale with higher score indicating easier comprehension. GL – Grade level represented by Flesch-Kincaid Grade Level score which rates text on a U.S. grade school level.

$SD = 15$) that indicates the probability of AD. The ASDS manual states that the purposes of the scale are fourfold: (a) to aid in the identification of persons who have AD, (b) to document behavioral progress, (c) to formulate target goals for Individualized Educational Programs (IEP), and (d) for use in research. Raters can be general education teachers, special education teachers, paraprofessionals, or parents; an appropriate rater should have two weeks of sustained contact with the individual being rated and should know the examinee well. The ASDS takes approximately 10–15 minutes to complete.

Test Construction and Standardization

Authors selected the 50 ASDS items based on review of diagnostic manuals, such as the DSM-IV, and a review of the literature on AD. The ASDS was standardized and normed using a small sample of 115 individuals diagnosed with AD (83% male) who ranged in age from 5 to 18 ($M = 10.42$, $SD = 3.44$) from 21 states across the United States. The standardization sample was recruited through two sources: (a) professionals in school districts (e.g., teachers, psychologists) who were asked to complete the ASDS on students previously diagnosed with AD, and (b) parents of children with AD through mailings and meetings where the test authors were invited to speak. Authors did not establish independent diagnosis for the standardization sample. Subtest and ASQ scores were created using cumulative frequency tables collapsed across age groups and gender, due to the absence of age and gender differences observed in the standardization sample.

Reliability, Floor, Ceiling, Item Gradient, and Readability

Authors provide evidence for internal consistency reliability and interrater reliability (see Table II). Within the standardization sample, Cronbach's coefficient alpha for the total score is .83, with subscale score coefficient alphas ranging from .64 (Cognitive) to .83 (Social) with a median internal consistency reliability value of .72. Parent and teacher interrater reliability is reported to be .93 for the ASQ for 14 pairs of raters. No information is provided for temporal stability (i.e., test–retest) reliability. Test authors recommend only the interpretation of the ASQ in decision making due to the unreliability of the ASDS subscales. The ASQ total score ranges from 30 to 135 with a maximum item gradient of 1/5 of a standard deviation, i.e., single raw score point equal to three standard score points. Therefore, the ASQ total score shows adequate floor, ceiling, and item gradients for the standardization

Table II. Reliability of Third-Party Ratings for Asperger's Disorder

	Internal Consistency [a]	Test-retest [b]	Interrater Reliability [b]
ASDS	.83/.64 – .83 Mdn = .72 (n = 115)	NR	.93 / NR (n = 14)[c]
ASSQ [d]	NR	.94 / NA (n = 65) [e] .96 / NA (n = 86)[f]	.77 / NA (n = 20)[c]
CAST [d]	NR	NR	NR
GADS	.87/.70–.81 Mdn = .77 (n = 360)	.93 / .71–.77 Mdn = .76 (n = 10)[e]	.89/.72–.84 Mdn = .82 (n = 16) [c, g]
KADI [d]	.93/NA (n = 130)	.98 / NA (n = 25)[h]	90% agreement (n = 19 pairs)[i]

Note. NA – Not applicable. NR – Not reported. Mdn – Median. ASDS – Asperger Syndrome Diagnostic Scale; ASSQ – Autism Spectrum Screening Questionnaire; CAST – Childhood Asperger Syndrome Test; GADS – Gilliam Asperger's Disorder Scale; KADI – Krug Asperger's Disorder Index.

[a] Total test score/Range of values for subtest scores. Statistics reported are Cronbach's coefficient alpha for individuals with Asperger's Disorder.

[b] Total test score/Range of values for subtest scores. Test-retest interval was two weeks for the ASSQ, ASDS, and KADI.

[c] Parent - teacher agreement.

[d] The ASSQ, CAST, and KADI do not contain subscales; therefore, subtest values are not applicable (NA).

[e] Teacher ratings.

[f] Parent ratings.

[g] Ten of 16 participants were diagnosed with Asperger's Disorder.

[h] Raters not identified in the examiner's manual.

[i] For KADI, interrater reliability calculation based upon percentage agreement between 19 pairs of raters. Agreement defined as standard score difference of less than 15 points.

sample. On average, ASDS subscale scores range from 2.6 to 15 with an average of 2.37 items per standard deviation across the range of possible scaled scores. Therefore, the ASDS shows an unacceptable average subtest ceiling and fails to meet Bracken's (1987) subtest item gradient criterion as raw scores are equivalent to almost 1/2 of a standard deviation, on average. The ASDS FRE score was 40.5 and F-KG score was 9.9.

Validity

Citing evidence for content validity, test authors report using rational item selection to create the list of 50 ASDS items, including review of DSM-IV and ICD-10 diagnostic criteria, literature review spanning 1975 to 1999 on AD, and review of Asperger's (1944) original research report. The test manual provides evidence of criterion validity as the ASQ total score correctly identified 85% of children across five classifications, including, AD, autism, behavioral disorder (BD), attention-deficit/hyperactivity disorder (ADHD), and learning disabled (LD). Evidence for construct validity includes: (a) statistically significant item-total score correlations for the ASDS items, (b) concurrent and discriminant validity with total scores and subscales of the Gilliam Autism Rating Scale (GARS), and (c) lack of statistically significant relationship between ASQ scores and age (r = .14) within the standardization sample. The ASDS and GARS total score correlation for 16 individuals is moderate (r = .46), but not statistically significant largely due to the small sample size.

Critique

Consistent with Goldstein (2002), the present author has serious concerns regarding the use of the ASDS, particularly the utility of the test to assist in differential diagnosis within the autism spectrum. Perhaps the most glaring weakness with the ASDS is the questionable standardization sample where independent diagnosis of AD was not determined. The authors also provide no evidence of cognitive functioning for the sample of individuals with autism in the validation study, an important omission as the Cognitive and Language subscales showed the greatest difference between the groups. It must be assumed that the autism group was cognitively impaired as up to 80% of children diagnosed with autism are also diagnosed with mental retardation. If the autism group showed cognitive impairment, the utility of the ASDS is diminished. Reliability data are also weak for the ASDS, as the ASQ fails to meet the .90 criterion for internal consistency and no temporal stability data are presented. The subtests for the ASDS also show ceilings that are too low, i.e., maximum scaled score is less than 16 for 4 of 5 subscales, and item gradients that are too steep for the ASDS subscales.

Autism Spectrum Screening Questionnaire (ASSQ)

Brief Description of Test

The ASSQ consists of 27 behavioral descriptions that are rated on a 3-point scale, to indicate if the rated "child stands out as different from other children of his/her age" (Ehlers *et al*., 1999, p. 139). The rater may endorse symptoms as not present ("0"), somewhat present ("1"), or definitely present ("2"). Items are summed to yield a total raw score that may range from 0–54. Items address problems in social interaction, communication, restricted and repetitive behavior, motor clumsiness and associated symptoms, such as the presence of motor and vocal tics. The ASSQ takes approximately 10 minutes to complete and is designed as a screening instrument to identify children who require more comprehensive evaluation to determine the presence of AD or high-functioning autism (HFA).

Test Construction and Standardization

ASSQ items were selected based on the authors' clinical experience with autism and "review of pertinent literature" (Ehlers *et al*., 1999, p. 130) with citations ranging from Asperger's (1944) original description to 1989. The pool of items was selected to reflect symptoms characteristic of AD in children 7–16 years of age. Drafts of the scale were reviewed by special education teachers in Sweden, and items were dropped or revised if ambiguous or misunderstood by teachers. The ASSQ was originally created and used in Ehlers and Gillberg's (1993) epidemiological investigation and reliability and validity established with a clinical population that consisted of two samples: (a) a sample of 110, 6- to 17-year-old children referred to a clinic setting with diagnoses of various autism spectrum disorders (ASD), attention-deficit and disruptive behavior disorders (ADD/BD), and learning disorders (LD), and (b) a validation sample of 34 children diagnosed with AD (Ehlers *et al*., 1999). Based upon findings from the clinical sample, parent scores of 19 and teacher scores of 22 are recommended cut-offs when deciding to refer for further evaluation.

Reliability, Floor, Ceiling, Item Gradient, and Readability

Within the epidemiological sample, Ehlers and Gillberg (1993) reported a test-retest reliability coefficient of .90 ($n = 139$) for ASSQ total scores among teacher ratings across an 8-month interval (see Table II). Ehlers and Gillberg also reported an interrater reliability coefficient of .79 ($n = 139$) for teacher–teacher agreement. Teacher ratings were conducted with typical children (Ehlers & Gillberg, 1993). In the clinical sample, Ehlers *et al*., (1999) reported a test–retest reliability coefficient of .94 ($n = 65$) for ASSQ total scores among teacher ratings and .96 ($n = 86$) for parent ratings for a 2-week interval. Authors do not report the number or percentage of children rated who were diagnosed with an ASD. Ehlers *et al*. also reported an interrater reliability coefficient of .77 ($n = 20$) for parent–teacher agreement for children diagnosed with an ASD. For the larger clinical sample, parent–teacher agreement was .66 ($n = 105$). The ASSQ has not been normed; therefore, floor, ceiling, and item gradient information is not available for the scale. The ASSQ has a FRE readability score of 46.3 and F-KG readability score of 8.7.

Validity

Ehlers *et al*. (1999) provide a variety of data regarding concurrent and divergent validity of the ASSQ (see Table III). Mean scores for the ASSQ significantly differed between the ASD, ADHD/BD, and LD diagnostic groups for both parent and teacher reports, while this was not the case for two non-ASD rating scales, the Rutter and Conners. Diagnosis for each group was determined via clinical case conference. Mean ASSQ scores for the validation sample of 34 children diagnosed with AD per Gillberg and Gillberg's (1989) criteria were described as similar to the original ASD sample, although no data were reported for mean scores from the original ASD sample. The ASSQ showed strong relationships with the Rutter and Conners' rating scales (rs = .58 and .77 for parent ratings, rs = .70 and .77 for teacher ratings) indicating statistically significant overlap that exceeds frequently cited standards (i.e., correlation between .30 and .70) published by Streiner (1993). The ASSQ has shown good specificity in correctly identifying non-AD cases and variable sensitivity for correctly detecting AD cases for both parent and teacher forms (Ehlers *et al*., 1999; see Table IV).

Critique

The authors of the ASSQ provide strong psychometric evidence for the scale's reliability with the notable omission of internal consistency reliability, which would seem to be easily accomplished with the authors' available data. Inter-rater consistency and temporal stability reliability data meet or exceed

Table III. Validity of Third-Party Ratings for Asperger's Disorder

	Content	Criterion	Construct
ASDS	DSM-IV; ICD-10; Asperger's 1944 article; Literature review spanning 1975–1999	Differentiates between AD, autism, LD, ADHD, behavior disorders. Correlates with GARS	Scores not dependent on age; Scale items correlate with total
ASSQ	Literature review spanning Asperger's 1944 article to 1989	Differentiates between ASD, attention/behavior disorders, LD Correlates moderately with Connors/Rutter's scales	None reported
CAST	DSM-IV; ICD-10; ASSQ items, PDD-Q items	Differentiates between ASD and typical children in pilot and community samples	None reported
GADS	DSM-IV-TR; ICD-10; Literature review. Other measures of AD reviewed, e.g., ASSQ	Differentiates between AD, autism, other disability groups, and non-disabled peers. Correlates with GARS	Scores not dependent on age or gender; Items correlate with total GADS score
KADI	Pool of items from ASIEP and other measures Literature review, e.g., Wing's (1981) description of AD	Differentiates between AD, high-functioning autism, and typical peers. Sensitivity, specificity, and positive predictive validity	Scores not dependent on age or gender; Items correlate with total KADI score

Note: AD – Asperger's Disorder; ASD – Autism Spectrum Disorder; ASDS – Asperger Syndrome Diagnostic Scale; ASSQ – Autism Spectrum Screening Questionnaire; CAST – Childhood Asperger Syndrome Test; GADS – Gilliam Asperger's Disorder Scale; GARS – Gilliam Autism Rating Scale; KADI – Krug Asperger's Disorder Index; LD – Learning Disability.

Table IV. Validity of Third-Party Ratings for Asperger's Disorder at Recommended Cut-off Scores

	% Correct classification[a]	Sensitivity	Specificity	PPV[b]
ASDS	85%	NR	NR	NR
ASSQ–Parent	NR	.62 – .82[c]	.90	NR
ASSQ–Teacher	NR	.65 – .70[c]	.91	NR
CAST	NR	.88	.98	.64
GADS	83%	NR	NR	NR
KADI	90%	.78	.94	.83

Note: NR – Not reported. ASDS – Asperger Syndrome Diagnostic Scale; ASSQ – Autism Spectrum Screening Questionnaire; CAST – Childhood Asperger Syndrome Test; GADS – Gilliam Asperger's Disorder Scale; KADI – Krug Asperger's Disorder Index.

[a] Discrimination between Asperger's Disorder and no disorder or Asperger's Disorder and other diagnostic groups.

[b] Positive Predictive Value.

[c] Ranges represent two samples reported for the ASSQ in Ehlers *et al.,* (1999).

acceptable standards. Inter-rater reliability appears to be quite strong. The scale has also been subjected to validation across three samples, a community-based sample, a general clinical sample, and clinical sample of children with AD. As a screener, the ASSQ shows adequate specificity but poor sensitivity for both the parent and teacher forms. No positive predictive validity values have been reported to date.

Childhood Asperger Syndrome Test (CAST)

Brief Description of Test

The CAST was designed specifically to screen school-age populations for behavioral symptoms indicative of AD (see Table I). The CAST is a 37-item parent-rating scale of behavioral indicators of AD scored as either present or absent. Of the 37 items on the CAST, 31 are summed to yield an overall score with 6 items sampling general development, which do not contribute to the total CAST score. Authors report that a CAST cut-off score of 15 or greater indicates the need for further evaluation for AD.

Test Construction and Standardization

The authors indicate that CAST items were selected from behavioral descriptions found in ICD-10 and DSM-IV diagnostic manuals relevant to the core features of ASDs (i.e., social and communication impairment, repetitive and/or stereotyped behaviors) as well as items from the ASSQ and the Pervasive

Developmental Disorders Questionnaire (PDD-Q). The CAST was piloted with a sample of 13 children with AD and 37 typically developing children.

Reliability, Floor, Ceiling, Item Gradient, and Readability

No information about reliability of the CAST has been published to date. The CAST has not been normed; therefore, floor, ceiling, and item gradient information is not available for the scale. The CAST FRE score was 74.9 and F-KG score was 4.1.

Validity

Parents of 50 children completed the CAST, 13 children with a prior diagnosis of AD or ASD and 37 typical children. Authors initially refer to the clinical sample as diagnosed with AD, but subsequently refer to the group as an AS/ASD group; therefore, it is not clear if the original sample of 13 children were diagnosed with AD only or included children diagnosed with other ASDs. Parents produced average CAST scores that differed significantly between the AD/ASD ($M = 21.08$, $SD = 5.51$) and typical ($M = 4.73$, $SD = 3.57$) samples. Chi-square analyses were conducted in the original sample with parents endorsing 27 of 31 items more frequently in the clinical group vs. the typical group. After pilot data were collected, the CAST was subjected to a larger validation study involving 1,150 children ranging in age from 4 to 11 who were enrolled in mainstream education. Of the 1,150 parents approached for participation, 199 (17.3%) parents responded, of these 199 a total of 139 children were evaluated for the presence of an ASD. Using a cut-off score of 15, the CAST yielded a sensitivity of .88 (7/8 cases), and a positive predictive value of .64 for the presence of an ASD diagnosis. The CAST yielded a specificity value of .98 at the 15-point cut-off score.

Critique

To date, no reliability data have been published for the CAST; therefore, the reliability criteria cannot be evaluated. The CAST validation study holds a number of desirable features including evaluation of the predictive validity of the CAST and the use of "gold standard" diagnostic instruments (e.g., Autism Diagnostic Observation Schedule (ADOS)) to establish psychiatric diagnosis after the screen. The CAST shows strong sensitivity and specificity in discriminating between AD and non-AD; however, the CAST holds poor positive predictive validity. Despite the use of widely accepted "gold standards" in the population-based study, it is not clear if the original sample included only children diagnosed with AD or a group of children diagnosed with AD and other ASDs.

Gilliam Asperger's Disorder Scale (GADS)

Brief Description of Test

The GADS is a 32-item norm-referenced rating scale that requires the respondent to indicate the frequency of behaviors indicative of AD across four subscales: social interaction, restricted patterns of behavior, cognitive patterns, and pragmatic skills. Raw scores are summed within the four domains and yield scaled scores ($M = 10$, $SD = 3$) and percentile ranks for subtests. Subtest scaled scores are summed for the entire scale to yield an Asperger's Disorder Quotient (ADQ), which is a standard score ($M = 100$, $SD = 15$) that indicates the probability of AD. The GADS also includes a Parent Interview Form to document the absence of clinical delays in language and cognitive development, adaptive behavior and curiosity about the environment, which are necessary for DSM-IV-TR diagnosis of AD. The Parent Interview Form items are not summed and do not contribute to the ADQ. The GADS manual states that the purposes of the scale are five-fold: (a) identify individuals with AD, (b) to assess persons who show unique behavioral features, (c) to document behavioral progress, (d) to target goals for IEPs, and (e) for use in research. Raters can be teachers, teacher's aides, parents, psychologists or psychological associates who have had at least 2 weeks of sustained contact with the individual being rated. The GADS manual states that most raters can complete the scale in about 5–10 minutes.

Test Construction and Standardization

Authors selected the 32 GADS items based on review of the DSM-IV-TR and ICD-10, a review of the literature on AD, and review of instruments designed to assess AD, including the ASSQ. An original pool of 70 experimental items was examined and reduced to 32 items. The GADS manual does not provide information about how the 38 items were discarded or how the 32 items were grouped into subscales. The manual indicates that some type of data analysis guided the final selection and grouping of the items, but information describing this process

is not provided in sufficient detail to evaluate. The GADS was standardized and normed using a sample of 371 individuals diagnosed with AD (85% male) who ranged in age from 3 to 22 ($M = 10$, $SD = 4$) from 46 states across the United States and countries outside of the United States. The standardization sample was recruited through two sources: (a) school professionals, who were asked to complete the GADS on students previously diagnosed with AD, and (b) parents of children with AD, who were contacted via the Internet. As with the ASDS, the test author did not establish independent diagnosis of AD for the GADS standardization sample. Subtest and ADQ scores were created using cumulative frequency tables due to the absence of age and gender differences observed in the standardization sample.

Reliability, Floor, Ceiling, Item Gradient, and Readability

Within the AD standardization sample ($n = 360$), Cronbach's coefficient alpha for the GADS total score is .87, with subscale score coefficient alphas ranging from .70 (Restricted Patterns of Behavior) to .81 (Cognitive Patterns) with a median subtest internal consistency reliability of .77. Temporal stability reliability is .93 for the ADQ for 10 teachers rating students with AD over a 2-week interval. Test–retest reliability ranged from .71 (Restricted Patterns of Behavior) to .77 (Pragmatic Skills) for GADS subscales. Parent and teacher interrater reliability is reported to be .89 for the ADQ for 16 children, 10 of whom were children with AD. The ADQ total score ranges from 40 to 132 with a maximum item gradient of 1/8 of a standard deviation, i.e., single scaled score point equal to two ADQ standard score points. Therefore, the ADQ total score shows adequate ceiling and item gradients for the standardization sample. On average, GADS subscale scores range from 1 to 14.75 with an average of 5.10 items per standard deviation across the range of subtest scaled scores. Therefore, the GADS shows adequate average item gradients at the subtest level as items are equivalent to approximately 1/5 of a standard deviation. The GADS yielded a FRE readability score of 55.7 and a F-KG readability score of 8.1.

Validity

The test author established content validity for the GADS by referring to the DSM-IV-TR and ICD-10 diagnostic criteria when creating test items. In 50 children referred to a university clinic, the GADS showed moderate to strong relationship with like subscales and total score of the GARS. The GADS discriminated between AD and a group of children diagnosed with autism and other disabilities, such as ADHD, LD, and mental retardation. Construct validity is addressed by the author documenting: (a) strong item-subscale correlations, (b) no differences between males and females on GADS scores, (c) lack of statistically significant correlation between age and GADS scores, with the exception of the Restricted Patterns of Behavior subscale, and (d) mean differences on the GADS between AD, autism, other disability groups, and non-disabled groups.

Critique

The GADS has the largest standardization group and offers some evidence that the AD standardization sample ($N = 371$) is representative of the larger AD population as evidenced by age-appropriate cognitive functioning in 33 of 371 individuals. Similar to the ASDS, the GADS ADQ does not meet the internal consistency reliability criterion of .90 for use with individuals with AD and the median subtest internal consistency falls below the criterion of .80. The average subscale ceiling (14.75) also falls short of the criterion of 16. A questionable group was used to create the norms for the GADS, as diagnoses were not verified by the test author.

Krug Asperger's Disorder Index (KADI)

Brief Description of Test

The KADI is a 32-item norm-referenced rating scale that requires the respondent to indicate the presence or absence of behaviors indicative of AD. Raw scores are weighted and summed to yield a KADI total standard score ($M = 100$, $SD = 15$) that indicates the likelihood for a diagnosis of AD with higher scores indicating greater likelihood of AD. The KADI consists of two groups of items, a subset of 11 items which are used as an initial screen for AD and the entire set of 32 items that contribute to the KADI total score. If the 11-item score does not exceed 18 raw score points, the rater is instructed to refrain from completing the remaining items. The KADI also consists of two forms, an Elementary form appropriate for ages 6–11 and a Secondary form appropriate for ages 12–21. Minor wording changes contained in the final five items constitute the

difference in the KADI forms. The KADI manual states that the purposes of the scale are threefold: (a) to identify individuals who have AD, (b) to target goals for intervention to be included in a student's IEP, and (c) for use in research. Authors identify an appropriate rater as someone who can read at the sixth-grade level who has regular and daily contact with the individual for at least a few weeks. The KADI manual does not provide a time estimate for completing the rating scale; however, parents who have completed the scale in The University of Georgia's School Psychology Clinic have done so in roughly 5–10 minutes.

Test Construction and Standardization

Authors generated an original pool of 106 items that was reduced to 32 based upon the items' ability to discriminate between AD and typical children, or AD and children with HFA. The KADI was standardized on a sample of 486 individuals, 130 diagnosed with AD, 162 diagnosed with autism, and 194 individuals described as "normal." Normative scores were calculated using the sample of 130 individuals with AD who were recruited from 32 states and 10 countries outside of the United States. Individuals with AD ranged in age from 6–0 to 21–11 with no mean or standard deviation reported in the KADI manual. The AD and autism samples were recruited through two sources: (a) mailing lists from various associations, centers, and institutes with interests in autism, such as the Autism Research Institute, and (b) professional conferences coordinated and conducted by Dr. Steve Edelson, an expert in autism and auditory integration training. The sample of typical individuals was recruited by "asking parents, teachers, and friends to use the KADI to rate one person between the ages of 6–0 and 21–11 with no previous diagnosis" (Krug & Arick, 2003, p. 14). Data were collected through surveys mailed to individuals that contained the original 106 items and demographic information. Based on the examiner's manual, authors did not establish independent diagnosis for the normative sample of individuals with AD or the sample of individuals with autism, but rather relied on diagnosis as reported by others.

Of the original 106 items, the 32 items retained in the final KADI were those that discriminated between the "normal" sample and the AD sample, i.e., the 11 screening items, and/or discriminated between the AD sample and the autism sample. The 32 items were weighted according to strength of relationship between item response and AD or autism diagnosis; items weights range from 0 – 4 for the KADI total score. For example, the 12 items that receive the highest weighted score of "4" are those with phi coefficients between .573 and .201 ($p < .003$). Two items with a weighted score of "0" were retained due to the items' ability to discriminate between AD and "normals" but not between AD and autism.

Reliability, Floor, Ceiling, Item Gradient, and Readability

Authors provide evidence for internal consistency reliability, temporal stability, and interrater reliability. Within the standardization sample, Cronbach's coefficient alpha for the KADI total score is .93, with temporal stability of .98 over a 2-week interval. Percent agreement of 90% is reported for 19 pairs of raters for individuals with AD, with agreement defined as standard scores falling within one standard deviation of each other. The KADI total score ranges from 60 to 129 with a maximum item gradient of 1/15 of a standard deviation, i.e., single raw score point equal to one standard score point. The KADI total score shows adequate item gradients for the standardization sample with the ceiling falling below the established criterion of two standard deviations above the normative sample mean. The KADI FRE score was 51.2 and F-KG score was 8.2, slightly higher than the authors' requirements that the rater read at a 6^{th}-grade level.

Validity

The original set of 106 items were selected from previously published rating scales, such as the Autism Screening Instrument for Educational Planning (Krug, Arick, & Almond, 1993) and others, review of Wing's (1981) description of AD, and Klin and Volkmar's (1995) review of AD. Authors provide evidence of the KADI's concurrent validity in the form of specificity, sensitivity, and positive predictive power (PPV) for discrimination between AD and typical samples as well as AD and autism samples. The KADI shows sensitivity of .78, specificity of .94, and PPV of .83 within the standardization sample of 486 individuals. Mean scores also significantly differ between AD, autism, and "normal" groups in the standardization sample. Construct validity for the KADI includes high item-total test correlations, which are not reported in the KADI manual.

Critique

The KADI presents the strongest set of reliability data among all five measures, although interrater reliability was calculated by percent agreement vs. correlation, which may inflate the KADI's reliability in this area. Similar to the ASDS and GADS, test authors did not confirm diagnosis of AD and autism. The test authors also do not provide data relevant to the cognitive functioning of the autism contrast group, which is of particular importance when differential diagnosis of AD and HFA is required. The vast majority of raters in the AD normative sample were relatives (94% of the sample), not teachers; however, the manual indicates that teachers are appropriate raters. Therefore, the information reported in the KADI manual pertains almost exclusively to ratings made by parents or other relatives (Table V).

SUMMARY AND CONCLUSIONS

Diagnosis of AD, especially the differentiation between AD and HFA, is a complicated venture due to a variety of factors, such as differing diagnostic systems and changing sets of criteria over time. Howlin (2000) summarized seven sets of diagnostic criteria for AD ranging from Asperger's (1944) original description to DSM-IV (1994) highlighting areas of similarity, difference, and omission. Given the range of definitions offered for AD, diagnosis is limited in a real sense by the absence of clear and satisfying criteria. Diagnostic criteria for AD have preceded scientific validation of AD as distinct from HFA, thereby putting the proverbial "cart before the horse" (Howlin, 2000, p. 120).

One significant reason that satisfying diagnostic criteria do not exist for AD is that no clinical, neuropsychological, or behavioral indicator reliably discriminates between children with AD and those with HFA. The search for "phenotypic" differences between AD and HFA have produced contradictory findings in the literature. For example, the verbal vs. non-verbal cognitive advantage for individuals with AD versus HFA documented by Klin, Volkmar, Sparrow, Cicchetti, and Rourke (1995) has been disputed by others (e.g., Manjiviona & Prior, 1999). Due to this state of affairs, it is not surprising that the rating scales reviewed showed significant limitations; however, *any* instrument that claims to diagnose AD must be able to discriminate between AD and HFA given the considerable overlap of diagnostic and associated features.

For the ASDS, GADS, and KADI, authors report the scales' ability to differentiate between AD and autism without reporting the cognitive functioning of the autism sample. In this case, matching on cognitive functioning is important for differential diagnosis otherwise differences found between groups become tautological as the test is merely confirming that groups are different in terms of cognitive and related language functioning. For example, the ASDS Cognitive and Language scales showed the greatest differentiation between AD and autism groups, which may merely reiterate the presence of a pre-existing difference in cognitive skills.

Commercially available instrumentation (i.e., ASDS, GADS, KADI) also shares consistent limitations regarding standardization and norming procedures. All ratings have been normed via mailing and survey methods without independent confirmation of diagnosis of AD; therefore, test users have no assurances that the normative sample consists only of individuals diagnosed with AD. Assuming that all survey respondents are, in fact, rating an individual

Table V. Summary of Three Commercial Instruments in Relationship to Bracken's (1987) Psychometric Criteria

Bracken's Psychometric Criteria (Standard)	ASDS	GADS	KADI
1. Total Test Internal Consistency (≥.90)	No	No	Yes
2. Median Subtest Internal Consistency (≥.80)	No	No	N/A
3. Total Test Temporal Stability (≥.90)	No	Yes	Yes
4. Total Test Ceiling (≥ 2 SDs above normative mean)	Yes	Yes	No
5. Average Subtest Ceilings (≥ 2 SDs above normative mean)	No	No	N/A
6. Average Subtest Floors (≥ 2 SDs below normative mean)	Yes	Yes	N/A
7. Average Subtest Item Gradient (average subtest item ≤ 1/3 SD for subtests)	No	Yes	N/A
8. Presence of Validity Evidence	Yes	Yes	Yes

Note: N/A – Not applicable. ASDS – Asperger Syndrome Diagnostic Scale; GADS – Gilliam Asperger's Disorder Scale; KADI – Krug Asperger's Disorder Index; SD – Standard Deviation.

with AD, one has no idea what definition of AD is being used to establish the diagnosis, and these vary widely (Howlin, 2000).

In light of the limitations described above and review of all data reported, the KADI showed the strongest psychometric properties and most thorough item selection among the three published measures; the ASDS consistently showed the weakest psychometric properties for the group. Both research measures appear to hold promise, with the ASSQ showing sound reliability and less convincing validity and the CAST showing sound predictive validity in the absence of published reliability data. All rating scales fall short of the standards set forth by Bracken (1987) and should be used with caution when evaluating the presence of AD and differentially diagnosing AD vs. HFA. Test authors appropriately recommend that their ratings scales be used as part of a larger evaluation when diagnosing AD vs. other possible conditions; results from the present review support such a cautious use. Future study comparing the utility of the rating scales in distinguishing between carefully diagnosed samples of children with AD vs. HFA would be helpful in determining the most appropriate measure to assist in differential diagnosis.

REFERENCES

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

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

Asperger, H. (1944). Die Autistischen Psychopathen im Kindesalter. *Archiv für Psychiatrie und Nervenkrankheiten*, *117*, 76–136.

Attwood, T. (1998). *Asperger's syndrome: A guide for parents and professionals*. London: Jessica Kingsley Publishers.

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

Bracken, B. A. (1987). Limitations of preschool instruments and standards for minimal levels of technical adequacy. *Journal of Psychoeducational Assessment*, *4*, 313–326.

Bracken, B. A., Keith, L. K., & Walker, K. C. (1994). Assessment of preschool behavior and social-emotional functioning: A review of thirteen third-party instruments. *Assessment in Rehabilitation and Exceptionality*, *4*, 331–346.

Campbell, J. M. (1998). Review of the Peabody Picture Vocabulary Test, Third Edition. *Journal of Psychoeducational Assessment*, *16*, 334–338.

Campbell, J. M., & Morgan, S. B. (1998). Asperger's Disorder. In L. Phelps (Ed.), *Health-related disorders in children and adolescents* (pp. 68–73). Washington, DC: American Psychological Association.

Ehlers, S., & Gillberg, C. (1993). The epidemiology of Asperger syndrome: A total population study. *Journal of Child Psychology and Psychiatry*, *34*, 1327–1350.

Ehlers, S., Gillberg, C., & Wing, L. (1999). A screening questionnaire for Asperger syndrome and other high-functioning autism spectrum disorders in school age children. *Journal of Autism and Developmental Disorders*, *29*, 129–141.

Ghuman, J. K., Freund, L., Reiss, A., Serwint, J., & Folstein, S. (1998). Early detection of social interaction problems: Development of a social interaction instrument in young children. *Developmental and Behavioral Pediatrics*, *19*, 411–419.

Gilliam, J. E. (2001). *Gilliam Asperger's Disorder Scale*. Austin, Texas: Pro-Ed Inc.

Goldstein, S. (2002). Review of the Asperger Syndrome Diagnostic Scale. *Journal of Autism and Developmental Disorders*, *32*, 611–614.

Howlin, P. (2000). Assessment instruments for Asperger syndrome. *Child Psychology and Psychiatry Review*, *5*, 120–129.

Klin, A., & Volkmar, F. R. (1995). *Asperger syndrome: Guidelines for assessment, diagnosis, and intervention*. Pittsburgh, PA: Learning Disabilities Association of America.

Klin, A., Volkmar, F. R., & Sparrow, S. S. (2000). *Asperger syndrome*. 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*, 1127–1140.

Krug, D. A., & Arick, J. R. (2003). *Krug Asperger's Disorder Index*. Austin, Texas: Pro- Ed Inc.

Krug, D. A., Arick, J. R., & Almond, P. (1993). *Autism Screening Instrument for Educational Planning*. Austin: Pro-Ed Inc.

Manjiviona, J., & Prior, M. (1999). Neuropsychological profiles of children with Asperger syndrome and autism. *Autism*, *3*, 327–356.

Mayes, S. D., Calhoun, S. L., & Crites, D. L. (2001). Does DSM-IV Asperger's disorder exist? *Journal of Abnormal Child Psychology*, *29*, 263–271.

Myles, B. S., Bock, S. J., & Simpson, R. L. (2001). *Asperger Syndrome Diagnostic Scale*. Los Angeles, California: Western Psychological Services.

Scott, F. J., Baron-Cohen, S., Bolton, P., & Brayne, C. (2002). The CAST (Childhood Asperger Syndrome Test): Preliminary development of a UK screen for mainstream primary-school age children. *Autism*, *6*, 9–31.

Streiner, D. L. (1993). A checklist for evaluating the usefulness of rating scales. *Canadian Journal of Psychiatry*, *38*, 140–148.

Wing, L. (1981). Asperger's syndrome: A clinical account. *Psychological Medicine*, *11*, 115–129.

World Health Organization (1992). *International classification of diseases and related health problems* (10th Ed.). Author-Geneva, Switzerland.