Psychological Assessment
1994, Vol. 6, No. 4, 284-290

1040-3590/94/$3.00

Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology

Domenic V. Cicchetti

In the context of the development of prototypic assessment instruments in the areas of cognition, personality, and adaptive functioning, the issues of standardization, norming procedures, and the important psychometrics of test reliability and validity are evaluated critically. Criteria, guidelines, and simple rules of thumb are provided to assist the clinician faced with the challenge of choosing an appropriate test instrument for a given psychological assessment.

Clinicians are often faced with the critical challenge of choosing the most appropriate available test instrument for a given psychological assessment of a child, adolescent, or adult of a particular age, gender, and class of disability. It is the purpose of this report to provide some criteria, guidelines, or simple rules of thumb to aid in this complex scientific decision. As such, it draws upon my experience with issues of test development, standardization, norming procedures, and important psychometrics, namely, test reliability and validity. As I and my colleagues noted in an earlier publication, the major areas of psychological functioning, in the normal development of infants, children, adolescents, adults, and elderly people, include cognitive, academic, personality, and adaptive behaviors (Sparrow, Fletcher, & Cicchetti, 1985). As such, the major examples or applications discussed in this article derive primarily, although not exclusively, from these several areas of human functioning.

Standardization Procedures

Although numerous assessment instruments used in the behavioral (and medical) sciences are not standardized appropriately on relevant demographic variables, the importance of this critical process cannot be underestimated. The standardization of any test of intelligence needs to be based on systematic stratification on the following variables: age; gender; education, occupation, or both; geographic region; and urban versus rural place of residence. In this sense, the Wechsler Adult Intelligence Scale—Revised (WAIS-R; Wechsler, 1981) "represents the ultimate in adult norming of a Wechsler battery" (Kaufman, 1990, p. 76). The same procedure was also used successfully in the development of the survey and expanded editions of the revised Vineland Adaptive Behavior Scales (hereinafter the revised Vineland; Sparrow, Balla, & Cicchetti, 1984a, 1984b), in the Kaufman Assessment Battery for Children (K-ABC; Kaufman & Kaufman, 1983), and, most recently, in the development of the Kaufman Adolescent and Adult Intelligence Test (KAIT; Kaufman & Kaufman, 1993).

Correspondence concerning this article should be addressed to Domenic V. Cicchetti, Veterans Affairs Medical Center and Yale University, West Haven, Connecticut 06516.

Norming Procedures

The appropriate standardization of a given assessment instrument, as just described, renders it possible to develop national norms for the valid interpretation of the meaning of a given person's scores on the standardized test. Applying appropriate methods, raw test scores are converted into several primary derived scores: standard scores, national percentile ranks, and age equivalents. In addition to providing the resulting primary norms, some standardized tests provide a number of supplementary norms based on special groups for which the test instrument may also be useful, such as emotionally disturbed, mentally retarded, visually challenged, and hearing-impaired samples (Sparrow, Balla, & Cicchetti, 1984a). In distinguishing between standardization and norming procedures, it must be stressed that a norm refers to the average score (or, more generally, the performance) of a standardization sample, however the latter is defined.

Another important feature of the standardization and norming process of some recently developed assessment instruments is the use of overlap samples to provide comparative information for a given subject on one or more tests that might either be used to supplement the information derived from the instrument of focus or (as discussed later) to provide specific information on certain components of test validity. For example, it is often important for a clinical examiner, teacher, or parent to understand the extent to which the same child may compare in terms of both cognitive and adaptive levels of functioning. This direct comparison cannot be made in the usual circumstance if the cognitive and adaptive behavior instruments used to evaluate the child have been normed on different standardization samples. To obviate this problem, the Vineland standardization program included two overlap samples. The first consisted of 719 children randomly selected from the national standardization sample for the K-ABC (Kaufman & Kaufman, 1983). The age range was between 2 years, 6 months and 12 years, 11 months. This overlap sample was administered both the revised Vineland and the K-ABC. A second group of 2,018 children, aged between 2 years, 6 months and 18 years, 11 months, was drawn randomly from the revised Vineland standardization sample and administered both the Vineland and the revised Peabody Picture Vocabulary Test (PPVT-R; Dunn & Dunn, 1981).

Such overlapping standardization samples result in comparative norms that are very useful to clinicians, teachers, parents, and administrators. For a more comprehensive treatment of this topic, specifically of how information deriving from cognitive and adaptive behavior tests that were normed and standardized on the same overlap sample can facilitate the development of specific remedial programs, see Cicchetti and Sparrow (1990); Cicchetti, Sparrow, and Rourke (1991); and Sparrow and Cicchetti (1985, 1987, 1989).

Perhaps the single most valuable derived score, based on successful test standardization and norming procedures, is the development of a series of standard scores that can then be used to determine the level of functioning of a given individual of a given age in comparison with her or his standardization peers, in terms of a percentile rank. The usual procedure, although there are exceptions, is to express a given standard score on a normative sample mean of 100, with a standard deviation of ±15. Because standard scores are normally (or Gaussian) distributed, a score of 100 is at the 50th percentile, a score of 85 is at the 16th percentile, and a score of 130 is beyond the 95th percentile of functioning. This test information derives from well-standardized and well-normed assessment instruments and is always applicable at the level of total or overall indices of functioning, such as an adaptive behavior composite score (Sparrow, Balla, & Cicchetti, 1984a; Sparrow & Cicchetti, 1989); a Full Scale IQ (e.g., Kaufman & Kaufman, 1993; Wechsler, 1981); or the K-ABC Mental Processing Composite (Kaufman & Kaufman, 1983). One notable exception to the 100 ± a standard deviation of 15 pertains to the fourth edition of the Stanford–Binet Intelligence Scale (Thorndike, Hagen, & Sattler, 1986), which was normed on a mean of 100 and a standard deviation of 16.

Although specific Vineland domains (e.g., Communication, Daily Living Skills, Socialization, and Motor Skills) are also normed on standardization samples with means of 100 and standard deviations of ±15, the typical subtest score for intelligence or IQ tests is often based on a mean of 10 and a standard deviation of ±3. Thus, on the KAIT, one can expect that 99% of a normal sample would produce standard scores on a given subtest (e.g., auditory comprehension) between a range of 1 and 19 (Kaufman & Kaufman, 1993).

Before leaving the topic of standard scores, it is important to mention briefly the concept of confidence intervals, sometimes referred to as bands of error. A given standard score can be banded with a range of confidence intervals that have been constructed to take into account the standard error of measurement of the test instrument. The standard error of measurement defines that amount of test–retest variability that is expected to occur on the basis of the inherent imprecision of the assessment instrument itself. Typical bands of error, for a given well-normed standardized test, are usually reported at one or more of the following confidence intervals: 68%, 85%, 90%, 95%, and 99% (e.g., Kaufman & Kaufman, 1993, p. 77; Sparrow, Balla, & Cicchetti, 1984a, p. 21).

Additional ways of interpreting scores deriving from well-standardized tests often include age equivalents and descriptive categories. The specific procedure for producing age equivalent scores is first to plot mean raw scores for each age group in the standardization sample on arithmetic graph paper. The age equivalent score for any given raw score value is read from a smoothed curve that has been fitted through the plotted points. Such scores are available for the revised Vineland domain and subdomain scores, the various subscales of the WAIS–R, and both the K-ABC and the KAIT. Age equivalents have the distinct advantage that they are easily understood by persons unfamiliar with statistics.

However, when raw score distributions are very uneven or skewed, as tends to be true when they are based on age equivalents, they do not provide the type of representation that Kaufman and Kaufman (1993) referred to as "the full scale continuum, as is necessary for deriving scaled scores" (p. 76). This provides the rationale for developing descriptive categories. For example, maladaptive levels can be derived to denote the frequency of a given individual's maladaptive behavior in comparison with that of peers of the same age in a national standardization sample. The maladaptive levels (i.e., descriptive categories) and the corresponding percentile ranks used to classify a given raw score, as reported by Sparrow, Balla, & Cicchetti (1984a), are as follows: nonsignificant (50th percentile and below), intermediate (51st–84th), and significant (85th percentile or higher). In a somewhat analogous fashion, Kaufman and Kaufman (1993) provided, for the aforementioned KAIT, raw score; percentile equivalent; and descriptive categories of average, below average, lower extreme, mild deficit, moderate deficit, and severe deficit. These are given for specific age groupings, ranging between 11 years and 85 years and over.

Because it is possible for a test to be adequately normed and standardized and yet have undesirable psychometric properties (e.g., poor reliability and validity), it becomes important to discuss these issues as they relate to assessment instruments.

Test Reliability

Reliability can take many forms. These often include a measure of internal consistency that defines the extent to which items in a given test, domain, subdomain, or subtest hang together. It is measured by application of the familiar coefficient alpha or, when items are scored dichotomously, by the Kuder–Richardson (KR-20) formula (Cronbach, 1970). Other forms of reliability include the familiar test–retest and interexaminer reliability. Finally, the stability of a trait or behavior over time is measured in terms of a temporal reliability coefficient (e.g., Cicchetti & Tyrer, 1988; Tyrer, Strauss, & Cicchetti, 1983). Depending on the scale of measurement for a given item, appropriate reliability coefficients would include kappa (i.e., nominally scaled data), weighted kappa (i.e., ordinally scaled data), or the intraclass correlation coefficient (i.e., dimensionally scaled data; e.g., Fleiss, 1981). Mathematical relationships and, under certain specified conditions, the mathematical equivalencies between the kappa or weighted kappa statistic on the one hand and the intraclass correlation statistic on the other, have been shown (a) in the dichotomous case (by Fleiss, 1975) and (b) in the ordinal case (by Fleiss & Cohen, 1973). Concerning coefficient alpha (or KR-20 in the dichotomous case), two comments need to be made, the first conceptual, the second biostatistical.

Although some biostatisticians would regard interexaminer reliability as the most important type of reliability assessment,

one noted statistician held the contrasting view that coefficient alpha is to be preferred over both test–retest and interexaminer reliability. Specifically, Nunnally (1978) noted that, "if coefficient alpha is low for a test, a relatively high correlation between retests should not be taken as an indication of high reliability" (p. 234).

How valid is this argument, however? The internal consistency of items in a test can be very low even though the items do, in fact, hang together perfectly. Thus, items with low or high ceilings can produce identical scores (e.g., for odd versus even items within a given subtest or domain), but, of course, will correlate zero.

Alternatively, there is the cogent argument that very high levels of internal consistency merely inform that items hang together well at a particular point in time. That is to say, the same level of internal consistency for the same subjects some weeks later may be based on completely different responses on the same test items at the two different times. Providing that the ordering of responses (e.g., between odd and even items) remains the same at each testing, coefficient alpha will be high, but test–retest reliability will be low. Similarly, if subjects were evaluated independently by two examiners at the same point in time (separated by a time interval large enough to rule out memory effects), then it is possible for coefficient alpha to be high for each examiner's evaluation, despite a low level of interexaminer agreement. It is this type of reasoning that would force most biostatisticians to disagree with the arguments of Nunnally and to focus on measures of internal consistency within the broader context of test–retest and interexaminer reliability. One might legitimately ask where indeed the entire field of diagnostic assessment in the behavioral sciences would be if scientists used measures of internal consistency as the primary index of the reliability of major nosologic systems rather than appropriate measures of interexaminer agreement. It is the training of independent examiners, using well-defined, nonoverlapping criteria (e.g., from the third edition of the *Diagnostic and Statistical Manual of Mental Disorders; DSM-III;* American Psychiatric Association, 1980) that has produced high levels of agreement that have revolutionized the field of neuropsychologic and neuropsychiatric diagnosis (e.g., see Grove, Andreasen, McDonald-Scott, Keller, & Shapiro, 1981).

Whether the reliability of a given assessment instrument is expressed in terms of a coefficient alpha, test–retest, interexaminer, or temporal reliability coefficient, it is useful to develop guidelines to distinguish levels that are clinically meaningful from those that may not be. Taking into account the caveats concerning item ceiling and floor effects and the need to consider coefficient alphas in the broader context of other types of reliability assessments (e.g., interexaminer), the following guidelines were suggested by Cicchetti and Sparrow (1990): When the size of the coefficient alpha or other measure of internal consistency is below .70, the level of clinical significance is unacceptable; when it is between .70 and .79, the level of clinical significance is fair; when it is between .80 and .89, the level of clinical significance is good; and when it is .90 and above, the level of clinical significance is excellent. Cicchetti and Sparrow (1990) went on to state that

> correlations of .70 or higher are usually considered acceptable levels of internal consistency of items. For both our target age range (infancy through age 5 years), as well as for all age groups in the Vineland standardization sample (through 18 years 11 months), this criterion was always met, with results consistently at the upper end of the acceptable range (.85 or higher). Results for infancy through the preschool years were as follows: coefficient alphas ranged between .89 and .94 for Communication; between .86 and .92 for Daily Living Skills; between .82 and .94 for Socialization; between .74 and .95 for Motor Skills, and between .96 and .98 for the Vineland Adaptive Behavior Composite (ABC). We would thus conclude that Vineland items meet adequately the criterion of acceptable levels of internal consistency both at a domain and an ABC level. (pp. 178–179)

With respect to evaluating levels of kappa, weighted kappa, or the intraclass correlation statistic used for measuring intra- and interexaminer levels of agreement, a number of biostatisticians have developed guidelines for determining levels of practical, substantive, or clinical significance (e.g., Cicchetti & Sparrow, 1981; Fleiss, 1981; Landis & Koch, 1977). The guidelines developed by Cicchetti and Sparrow (1981) resemble closely those developed by Fleiss (1981) and also represented a simplified version of those introduced earlier by Landis and Koch (1977). The guidelines state that, when the reliability coefficient is below .40, the level of clinical significance is poor; when it is between .40 and .59, the level of clinical significance is fair; when it is between .60 and .74, the level of clinical significance is good; and when it is between .75 and 1.00, the level of clinical significance is excellent.

Before leaving the topic, it is important to mention that a number of reliability measurements of major assessment instruments in the field of intelligence testing (as well as in other fields) have been based on the standard Pearson product–moment correlation (r) rather than on the statistic of choice mentioned here, namely, the intraclass correlation coefficient (r_I). The problem with the product–moment correlation is that it measures similarity in the orderings of test scores made by independent evaluators. Thus, two independent examiners might be very far apart in the total IQ they attribute to the same group of adolescents. However, to the extent that their IQ rankings covary in the same order, the resulting correlation can range between very high and perfect. As noted by Kazdin (1982), "The correlation merely assesses the extent to which scores go together and not whether they are close to each other in absolute terms" (p. 58).

The aforementioned intraclass correlation coefficient, in contrast to the product–moment correlation, has the following desirable properties: (a) It can distinguish those paired assessments made by the same set of examiners from those made by different sets of examiners; (b) it distinguishes those sets of scores that are merely ranked in the same order from test to retest from those that are not only ranked in the same order but are in low, moderate, or complete agreement with each other; and (c) it corrects for the extent of test–retest (or interexaminer) agreement expected on the basis of chance alone (e.g., Bartko, 1966, 1974; Bartko & Carpenter, 1976; Cicchetti & Sparrow, 1981, 1990; Fleiss, 1981).

The question is, however, under what circumstances do the intraclass correlation coefficient and the product–moment correlation produce similar, dissimilar, or the same values? This question is easily answered. The product–moment correlation places the maximum limit on what the intraclass correlation

Table 1
Hypothetical Data for Comparison of Intraclass Correlation Coefficient and Pearson Product-Moment Correlation

Clinical applicant	Clinical neuropsychologist						
	A	B	C	D	E	F	G
1	10	10	9	8	7	6	5
2	9	9	8	7	6	5	4
3	8	8	7	6	5	4	3
4	7	7	6	5	4	3	2
5	6	6	5	4	3	2	1

coefficient can be. Thus, the intraclass correlation coefficient can be no higher than the product–moment correlation and will be lower than the product–moment correlation depending on the extent to which there is a systematic bias (or higher mean values) for one examiner's set of evaluations in relation to that of another.

Suppose that seven clinical psychologists each evaluate the same five internship applicants on a 10-category ordinal scale with respect to suitability for a particular program of predoctoral clinical training in neuropsychology. Suppose, in addition, that specific criteria (i.e., anchorage points) define these 10 categories as follows: 10 = outstanding candidate, top priority; 7 = equal in quality to the average intern selected at this institution; 4 = acceptable, but just barely; and 1 = completely unacceptable. Finally, assume that the scale points 9, 8, 6, 5, 3, and 2 denote, respectively, clinical evaluations that fit between the four anchorage points, with progressively more negative denotation. Let us say that the hypothetical data are those presented in Table 1.

The statistic of choice would be the intraclass correlation (Model II), which assumes the same set of five examiners throughout (e.g., Fleiss, 1981; see also Shrout & Fleiss, 1979). Results are presented in Table 2.

The point is obvious, that is, the product–moment correlation does not measure levels of agreement, but the intraclass correlation coefficient does just that. As the difference in paired scores increases systematically from 0 to 5 points, respectively, the intraclass correlation coefficient decreases from 1.00 to .83, .56, .36, .24, and .17. The product–moment correlation remains constant at 1.00.

In the next section, I focus on the important problem of determining the extent to which a clinical assessment instrument is valid or measures what it purports to measure. A much more articulate way of saying this (i.e., a way that eliminates the tautologic quality of the more traditional one just given) was provided by Kaufman and Kaufman (1993), who stated, simply, that "the validity of a test is defined as the degree to which it accomplishes what it was designed to do" (p. 84).

Table 2
Hypothetical Results in Comparison of Intraclass Correlation Coefficients and Pearson Product–Moment Correlations

Psychologist pairing	Number of scale points apart	Values of r	Corresponding values of r_I
AB	0	1.00	1.00
AC	1	1.00	.83
AD	2	1.00	.56
AE	3	1.00	.36
AF	4	1.00	.24
AG	5	1.00	.17

Test Validity

When clinical and research scientists collaborate on the challenging task of developing a new assessment instrument in psychology, and they have already carefully delineated the specific areas of interest (e.g., intelligence, adaptive behavior, personality development), the specific age range of focus (e.g., across the life span, adolescence, childhood), and whether the test is to be applied to specific disability groups (e.g., mentally retarded, learning disabled, or stroke samples), the arduous task of developing specific test items begins.

A major goal of test item development is to obtain as comprehensive a range of content coverage as will do justice to a full range of the meaning of the concept being measured (e.g., personality changes following left- and right-hemisphere stroke as in Nelson et al., 1993).

This "*content-related*" validity, as coined by Fitzpatrick (1983), derives from a number of major sources, namely, a comprehensive review of the relevant content literature; the test developers' clinical, educational, and research experiences; and pilot testing of large pools of items (e.g., Kaufman & Kaufman, 1993, in the development of the KAIT; Sparrow, Balla, & Cicchetti, 1984a, 1984b, 1985, in the development of the revised Vineland; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989, in the development of the Minnesota Multiphasic Personality Inventory-2 [MMPI-2]; and Millon, 1987, in the development of the Millon Clinical Multiaxial Inventory-II).

A given test item is constructed to produce responses on one of several possible scales of measurement: dichotomous, with two categories of response; polychotomous, with three or more nonordered (qualitative) response categories (e.g., Fleiss, 1981); continuous ordinal or dichotomous ordinal (Cicchetti, 1976; Cicchetti & Sparrow, 1981; Cicchetti, Volkmar et al., 1992); or continuous, interval, or dimensionally scaled data (Feinstein, 1987, pp. 24–25). Specific administration and scoring rules are provided, in the form of a manual, that can be used to train clinical examiners who plan to use the assessment instrument in the most reliable manner possible.

Once the instrument has been successfully field tested or pilot tested, with items retained, deleted, revised, or added to the final product, a second type of validity assessment is possible, namely, *face validity.* Do the items indeed look as though they measure what they are intended to measure? Drawing from the revised Vineland, it would be most unlikely that items measuring the receptive subdomain of Communication (i.e., what the individual understands) would ever be confused with either the expressive subdomain (i.e., what the individual says) or the written subdomain (i.e., what the individual reads and writes).

Discriminant validity is evidenced by the extent to which a relevant behavior or other test response is performed differentially by specifically selected samples in accordance with expec-

tations or hypothesized relationships among the selected groups. Several examples of this type of validity measurement were obtained using the aforementioned Vineland supplementary norm groups, namely, visually handicapped, hearing-impaired, and emotionally disturbed residential children.

As hypothesized, (a) emotionally disturbed children (i.e., by *DSM–III* criteria) showed their most serious deficits in the Vineland Socialization domain and manifested significantly more maladaptive behaviors than was true of their nonhandicapped peers in the standardization program, (b) visually handicapped children showed the most extensive deficits of any of the supplementary norm groups in overall adaptive behavior, and (c) hearing-impaired children evidenced their greatest deficits on the Vineland Communication domain (Cicchetti & Sparrow, 1990).

Kaufman and Kaufman (1993) used information from KAIT supplementary norms with a different objective in mind, namely, to provide what they refer to as *clinical validity* samples. By matching these clinical samples to controls on the demographic variables of age, gender, race or ethnicity, and educational status, the authors were able to provide some convincing preliminary evidence of the potential applicability of the KAIT "for assessing a variety of clinical and neurological problems" (p. 107). The clinical validity samples included people who were neurologically impaired (i.e., right-hemisphere impairment), clinically depressed, and reading disabled; as well as people with dementia of the Alzheimer's type. The authors are to be commended for expressing caution until the results of appropriate cross-validation samples are available. In a broader sense, the example presented illustrates the manner in which studies can be designed appropriately for testing the extent to which major cognitive assessment instruments may provide specific diagnostic information, in the form of normative data, that might be useful for applying the test to samples other than those on which it was originally nationally normed and standardized.

Concurrent validity refers to the extent to which a new assessment instrument correlates with an earlier instrument measuring the same or, more likely, a similar construct. In this particular area, it is not possible to generate useful rules of thumb concerning an ideal or minimally or maximally useful correlation value, as so much depends on what the new test purports to measure in relation to the old one. We know for sure that we would hope for a correlation of neither 1.00 nor 0. In the first case, the new test could be considered a veritable clone of the one with which it is being compared. In the second case, the construct validity of the very concept being measured would be called into question. Another important factor is the extent to which societal and cultural changes since the development of a new test may have necessitated major item changes, as in the case of what was considered adaptive behavior more than three decades ago (e.g., the Vineland Social Maturity Scale; Doll, 1935, 1965) in comparison with today. For example, it was considered adaptively appropriate for children to roam their neighborhoods unattended during Doll's era, but in many neighborhoods today this may be viewed as maladaptive (or potentially life threatening). Another factor is the extent to which earlier instruments may have been inappropriately standardized or normed, for example, the original Vineland Social Maturity Scale, normed on an unrepresentative sample from Vineland, New Jersey; or the original version of the MMPI (Hathaway & McKinley, 1940), normed on a sample based in Minnesota, in comparison with the more broad-based norms produced for the MMPI-2 (Butcher et al., 1989). Examples of concurrent validity of nationally normed adaptive behavior and IQ tests can be found in Sparrow and Cicchetti (1989) and in Kaufman and Kaufman (1993).

In terms of *factorial validity*, as noted by Sparrow, Balla, and Cicchetti (1984a, pp. 43–44) and Sparrow and Cicchetti (1989), two types of factor analyses were undertaken (principal component, principal factor) on the Vineland nationally normed and standardized sample of nonhandicapped U.S. subjects. In general, subdomains and their respective items loaded appropriately on their intended domains, that is to say, in general, the fit of subdomains (e.g., receptive, expressive, written language) into their intended domains (i.e., Communication) was quite successful. For example, for children 2–3 years of age, the written subdomain, as expected, did not correlate significantly with the factor labeled Communication. Consistent with this result, for children 8–9 years of age, the receptive domain did not fit into the Communication domain to any significant degree. This is also to be expected. Moreover, for both younger (aged 2 to 3 years) and older (aged 8 to 9 years) children, the fit of subdomains into their respective domains was highly significant. For example, the three subdomains interpersonal relationships, play and leisure time, and coping skills were the most highly correlated with the Socialization factor, "which is comprised precisely of these three subdomains" (Sparrow & Cicchetti, 1989, p. 212).

Kaufman and Kaufman (1993) argue convincingly that "factor structure is probably the most important evidence of a theory-based, multiscale test's construct validity" (pp. 90–95). Using both exploratory and confirmatory factor analyses on their KAIT data, Kaufman and Kaufman provided impressive results indicating that the two subtests defining the KAIT emerge consistently across age groups. Specifically, the two factors coincide very closely with the division of subtests into the two defined scales of the KAIT, namely, Fluid and Crystallized Scales of Adult Intelligence.

Before leaving this important section, one needs to focus on a special application of validity assessment that has been used with some well-known assessment instruments in psychology but derives originally from the field of medical diagnosis. In a number of (but not all) areas of medicine, it is possible to do confirmatory laboratory examinations that serve as the gold standard for the presence or absence of certain diseases. Such confirmatory evidence defines what might be referred to as *criterion validity*. Using this paradigm, one can compare the physician's test result (e.g., a diagnosis of bacteremia) with the results of a confirmatory result. The usual fourfold contingency table that is thereby generated is illustrated in Table 3. The four cells can then be identified, in terms of the examination's sensitivity, specificity, positive predictive value, and negative predictive value, as follows: Sensitivity, defined as $a/(a + c)$, refers to the extent to which the cases confirmed as positive by the laboratory test positive by the clinician (also referred to as true positive cases). Specificity, defined as $d/(b + d)$, refers to the extent to which the cases confirmed as negative by the laboratory test

Table 3
Hypothetical Fourfold Contingency Table

Physician exam result	Confirmatory lab result: Positive	Confirmatory lab result: Negative	Total
Positive (+)	(+ +) (a)	(+ −) (b)	(a + b)
Negative (−)	(− +) (c)	(− −) (d)	(c + d)
Total	(a + c)	(b + d)	*N*

negative by the physician (also referred to as true negative cases). Overall accuracy, defined as (a + d)/*N*, refers to the extent to which all tested cases (true positive plus true negative) produce the correct diagnosis. Positive predictive value, defined as a/(a + b), refers to the extent to which the test positive cases of the physician are confirmed positive by the laboratory. Negative predictive value, defined as d/(c + d), refers to the extent to which the test negative cases of the physician are confirmed negative by the laboratory. The false positive cases are defined by the (+ −) or the b cell, and the false negative cases are defined by (− +) or the c cell.

Although it is not yet possible to develop laboratory confirmed diagnoses of personality or mental disorders in the behavioral sciences, in recent years, some behavioral scientists have used the "best clinician diagnosis" (i.e., positive or negative) as a criterion, or gold standard, against which to compare the results of a normed and standardized test (i.e., positive or negative). Recent examples follow. Using a combined score of 10 or higher on the Beck Depression Inventory (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) and a simultaneous and independent *DSM-III* diagnosis of clinical depression, as the criterion, and established MMPI standard score cutoff points defining depression, we were able to show in a sample of outpatients the following: an MMPI (a) sensitivity of 78%, (b) specificity of 75%, (c) overall diagnostic accuracy of 77%, (d) positive predictive value of 93%, and (e) negative predictive value of 43% (Nelson & Cicchetti, 1991).

For a much more comprehensive approach to this problem, see the theory-driven work of Millon (1987), who succinctly summarized the need for such an approach:

> Diagnostic instruments are more useful when they are linked systematically to a comprehensive clinical theory. Unfortunately, as many have noted (Butcher, 1972), assessment techniques and personality theorizing have developed almost independently. As a result, few diagnostic measures have either been based on or have evolved from clinical theory. The MCMI-II is different. Each of its 22 clinical scales was constructed as an operational measure of a syndrome derived from a theory of personality and psychopathology (Millon, 1969, 1981). As such, the scales and profile of the MCMI-II measure theory-derived variables directly and quantifiably. Since these variables are anchored to a broad-based and systematic theory, they suggest specific patient diagnoses and clinical dynamics, as well as testable hypotheses about social history and current behaviors. (Millon, 1987, p. 3)

Summary

In summary, this article represents an attempt to highlight the need for well-normed and standardized test instruments in psychology in particular and in the behavioral sciences in general. As appropriate, guidelines, criteria, and rules of thumb have been provided to help clinicians arrive at a decision as to which among a myriad of available test instruments might be most appropriate for a given psychological assessment. It is not intended to be a comprehensive survey of the entire literature of appropriately normed and standardized assessment instruments, a colossal endeavor that is far beyond the stated objectives. Rather, I chose to focus mainly on known test instruments that have clearly defined and desirable psychometric properties and have very desirable norms that are based on appropriate standardization procedures. For comprehensive reviews of these basic issues in the field of intelligence testing, the interested reader is referred to Kaufman's (1990) scholarly, comprehensive, and insightful book *Assessing Adolescent and Adult Intelligence* and, more recently, to Kaufman and Kaufman's 1993 test manual for the KAIT. For a comprehensive review of other adaptive behavior scales, see Reschly (1987), Salvia and Ysseldyke (1988), and Sattler (1987).

References

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports, 19,* 3–11.

Bartko, J. J. (1974). Corrective note to: "The intraclass correlation coefficient as a measure of reliability." *Psychological Reports, 34,* 418.

Bartko, J. J., & Carpenter, W. T. (1976). On the methods and theory of reliability. *Journal of Nervous and Mental Disease, 163,* 307–317.

Beck, A., Ward, C., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives for Psychiatry, 4,* 561–571.

Butcher, J. N. (Ed.). (1972). *Objective personality assessment.* New York: Academic Press.

Butcher, J., Dahlstrom, W., Graham, J., Tellegen, A., & Kaemmer, B. (1989). *Minnesota Multiphasic Personality Inventory-2 (MMPI-2): Manual for administration and scoring.* Minneapolis: University of Minnesota Press.

Cicchetti, D. V. (1976). Assessing inter-rater reliability for rating scales: Resolving some basic issues. *British Journal of Psychiatry, 129,* 452–456.

Cicchetti, D. V., & Sparrow, S. S. (1981). Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency, 86,* 127–137.

Cicchetti, D. V., & Sparrow, S. S. (1990). Assessment of adaptive behavior in young children. In J. J. Johnson and J. Goldman (Eds.), *Developmental assessment in clinical child psychology: A handbook* (pp. 173–196). New York: Pergamon Press.

Cicchetti, D. V., Sparrow, S. S., & Rourke, B. P. (1991). Adaptive behavior profiles of psychologically disturbed and developmentally disabled persons. In J. L. Matson and J. Mulich (Eds.), *Handbook of mental retardation* (pp. 222–239). New York: Pergamon Press.

Cicchetti, D. V., & Tyrer, P. (1988). Reliability and validity of personality assessment. In P. J. Tyrer (Ed.), *Personality disorders: Diagnosis, management and course* (pp. 63–73). London: Butterworth Scientific.

Cicchetti, D. V., Volkmar, F., Sparrow, S. S., Cohen, D., Fermanian, J., & Rourke, B. P. (1992). Assessing the reliability of clinical scales when the data have both nominal and ordinal features: Proposed guidelines for neuropsychological assessments. *Journal of Clinical and Experimental Neuropsychology, 14,* 673–686.

Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd ed.). New York: Harper & Row.

Doll, E. A. (1935). A genetic scale of social maturity. *The American Journal of Orthopsychiatry, 5,* 180–188.

Doll, E. A. (1965). *The Vineland Social Maturity Scale.* Circle Pines, MN: American Guidance Service.

Dunn, L. M., & Dunn, L. (1981). *Manual for the Peabody Picture Vocabulary Test—Revised (PPVT-R).* Circle Pines, MN: American Guidance Service.

Feinstein, A. R. (1987). *Clinimetrics.* New Haven, CT: Yale University.

Fitzpatrick, A. R. (1983). The meaning of content validity. *Applied Psychological Measurement, 7,* 3–13.

Fleiss, J. L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics, 31,* 651–659.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: Wiley.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33,* 613–619.

Grove, W. M., Andreasen, N. C., McDonald-Scott, P., Keller, M. B., & Shapiro, R. W. (1981). Reliability studies of psychiatric diagnosis: Theory and practice. *Archives of General Psychiatry, 38,* 408–413.

Hathaway, S. R., & McKinley, J. C. (1940). A multiphasic personality schedule (Minnesota): I. Construction of the schedule. *Journal of Psychology, 10,* 249–254.

Kamphaus, R. W. (1993). *Clinical assessment of children's intelligence.* Boston: Allyn & Bacon.

Kaufman, A. S. (1990). *Assessing adolescent and adult intelligence.* Boston: Allyn & Bacon.

Kaufman, A. S., & Kaufman, N. L. (1983). *Kaufman Assessment Battery for Children (K-ABC) administration and scoring manual.* Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Kaufman, N. L. (1993). *Kaufman Adolescent and Adult Intelligence Test (KAIT) manual.* Circle Pines, MN: American Guidance Service.

Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings.* New York: Oxford University Press.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174.

Millon, T. (1969). *Modern psychopathology.* Philadelphia: W. B. Saunders.

Millon, T. (1981). *Disorders of personality: DSM-III, Axis II.* New York: Wiley.

Millon, T. (1987). *Millon Clinical Multiaxial Inventory-II: Manual for the MCMI-II.* Minneapolis: National Computer Systems, Inc.

Nelson, L. D., & Cicchetti, D. (1991). Validity of the MMPI Depression scale for outpatients. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 3,* 55–59.

Nelson, L., Cicchetti, D. V., Satz, P., Stern, S., Sowa, M., Metrushina, M., & Van Gorp, W. (1993). Emotional sequelae of stroke. *Neuropsychology, 7,* 553–560.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Reschly, D. J. (1987). *Adaptive behavior in classification and programming with students who are handicapped* (Monograph). Minneapolis, MN: Department of Education.

Salvia, J., & Ysseldyke, J. (1988). *Assessment in special and remedial education* (4th ed.). Boston: Houghton Mifflin.

Sattler, J. M. (1987). *Assessment of children's abilities* (3rd ed.). San Diego, CA: Author.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428.

Sparrow, S. S., Balla, D. A., & Cicchetti, D. V. (1984a). *The Vineland Adaptive Behavior Scales: A revision of the Vineland Social Maturity Scale by Edgar A. Doll. I. Survey form.* Circle Pines, MN: American Guidance Service.

Sparrow, S. S., Balla, D. A., & Cicchetti, D. V. (1984b). *The Vineland Adaptive Behavior Scales: A revision of the Vineland Social Maturity Scale by Edgar A. Doll. II. Expanded form.* Circle Pines, MN: American Guidance Service.

Sparrow, S. S., Balla, D. A., & Cicchetti, D. V. (1985). *The Vineland Adaptive Behavior Scales: A revision of the Vineland Social Maturity Scale by Edgar A. Doll. III. Classroom edition.* Circle Pines, MN: American Guidance Service.

Sparrow, S. S., & Cicchetti, D. V. (1985). Diagnostic uses of the Vineland Adaptive Behavior Scales. *Journal of Pediatric Psychology, 10,* 215–225.

Sparrow, S. S., & Cicchetti, D. V. (1987). Adaptive behavior and the psychologically disturbed child. *Journal of Special Education, 21,* 89–100.

Sparrow, S. S., & Cicchetti, D. V. (1989). The Vineland Adaptive Behavior Scales. In C. S. Newmark (Ed.), *Major psychological assessment instruments* (pp. 199–231). Boston: Allyn & Bacon.

Sparrow, S. S., Fletcher, J. M., & Cicchetti, D. V. (1985). Psychological assessment of children. In R. Michels, J. O. Cavenar, H. K. H. Brodie, A. M. Cooper, S. B. Guze, L. L. Judd, G. L. Klerman, & A. J. Solnit (Eds.), *Psychiatry* (Vol. 2, pp. 1–12). Philadelphia: Lippincott.

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986). *Stanford-Binet Intelligence Scale* (4th ed.). Chicago: Riverside.

Tyrer, P., Strauss, J., & Cicchetti, D. V. (1983). Temporal reliability of personality in psychiatric patients. *Psychological Medicine, 13,* 393–398.

Wechsler, D. (1981). *Manual for the Wechsler Adult Intelligence Scale—Revised (WAIS-R).* New York: Psychological Corporation.

Received January 5, 1994
Revision received May 16, 1994
Accepted May 18, 1994 ■