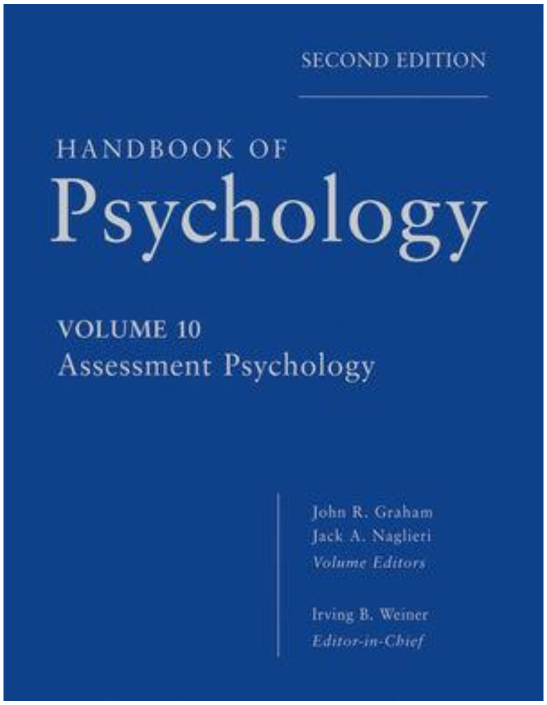

SECOND EDITION

HANDBOOK OF
Psychology

VOLUME 10
Assessment Psychology

John R. Graham
Jack A. Naglieri
Volume Editors

Irving B. Weiner
Editor-in-Chief

To cite this chapter:

Wasserman, J. D., & Bracken, B. A. (2013). Fundamental psychometric considerations in assessment. In I. B. Weiner, J. R. Graham, & J. A. Naglieri (Eds.), *Handbook of psychology. Assessment psychology* (2nd ed., Vol. 10, pp. 50-81). Hoboken, NJ: John Wiley & Sons.

To access this volume online, please visit:

http://online library.wiley.com

To purchase this volume from the publisher, please visit:

http://www.wiley.com/WileyCDA/WileyTitle/productCd-0470891270.html

CHAPTER 3

Fundamental Psychometric Considerations in Assessment

JOHN D. WASSERMAN AND BRUCE A. BRACKEN

"Whenever you can, count!" advised Sir Francis Galton, according to his biographer Karl Pearson (1924, p. 340), who reported that Galton seldom went for a walk or attended a lecture without counting something. The father of contemporary psychometrics, Galton is credited with applying the normal probability distribution to the study of individual differences and initiating the first large-scale efforts to measure physical, sensory, motor, and higher-order mental characteristics in his London Anthropometric Laboratories. Moreover, Galton is recognized for his discovery of the phenomena of regression, his conceptualization of the covariance between variables as a basis for understanding bivariate relations (with the product-moment correlation coefficient introduced by Pearson), and his execution of the first multivariate analyses (e.g., Stigler, 1999, 2010). Galton quantified everything from fingerprint characteristics, to variations in weather conditions, to the number of brush strokes taken by artists while he sat for portraits. At scientific meetings, he was known to count the number of times per minute that members of the audience fidgeted, computing an average and deducing that the frequency of fidgeting was inversely associated with level of audience interest in the presentation.

Of course, the challenge in contemporary assessment is to know what to measure, how to measure it, and when the measurements are meaningful. In a definition that still remains appropriate, Galton (1879) defined *psychometry* as "the art of imposing measurement and number upon operations of the mind" (p. 149). Derived from the Greek *psyche* (ψυχή, meaning "soul") and *metro* (μετρώ, meaning "measure"), psychometry may be best considered an evolving set of scientific rules for the development and application of psychological tests.

Construction of psychological tests is guided by psychometric theories in the midst of a paradigm shift. Classical test theory (CTT), epitomized by Gulliksen's (1950) *Theory of Mental Tests,* dominated psychological test development through the latter two-thirds of the 20th century. Item response theory (IRT), beginning with the work of Rasch (1960) and Lord and Novick's (1968) *Statistical Theories of Mental Test Scores,* is growing in influence and use, with calls by its advocates for a "velvet revolution" in psychological measurement (Borsboom, 2006b, p. 467). Embretson (2004) summarized the current status of the paradigm shift from CTT to IRT: "[A]t the end of the 20th century, the impact of IRT on ability testing was still limited. Only a few large-scale tests had applied IRT by the late 1990s. The majority of psychological tests still were based on classical test theory that was developed early in the 20th century" (p. 8).

This chapter describes the most salient psychometric characteristics of psychological tests, incorporating elements from both CTT and IRT. It provides guidelines for the evaluation of test technical adequacy. Although psychometricians frequently warn that such guidelines are oversimplified, we consider them to be rules of thumb that have practical value for test consumers using an applied handbook. The guidelines may be applied to a wide array of tests, including those in the domains of academic achievement, adaptive behavior, cognitive-intellectual abilities, neuropsychological functions, personality and psychopathology, and personnel selection. The guidelines are based in part on conceptual extensions of the *Standards for Educational and Psychological Testing* (1999; currently undergoing revision) and recommendations from such authorities as Anastasi and Urbina (1997; see also Urbina, 2004); Bracken (1987); Cattell (1986); Cohen (1992); Neuendorf (2002); Nunnally & Bernstein (1994); Salvia, Ysseldyke, & Bolt (2010); and Streiner (2003).

PSYCHOMETRIC THEORIES

The psychometric characteristics of mental tests are generally derived from one or both of the two leading theoretical approaches to test construction: CTT and IRT. Although it is common for psychometricians to contrast these two approaches and advocate for more contemporary techniques (e.g., Embretson, 1995; Embretson & Hershberger, 1999), most contemporary test developers in practicality use elements from both approaches in a complementary manner (e.g., Nunnally & Bernstein, 1994). For many challenges in test development, CTT and IRT findings may be largely interchangeable; Fan (1998) reported empirical findings indicating that person and item statistics derived from both theories are functionally comparable.

CLASSICAL TEST THEORY

CTT traces its origins to the procedures pioneered by Galton, Pearson, C. E. Spearman, and E. L. Thorndike, and is usually defined by Gulliksen's (1950) classic book. CTT has shaped contemporary investigations of test score reliability, validity, and fairness as well as the widespread use of statistical techniques such as factor analysis.

At its heart, CTT is based on the assumption that an obtained test score reflects both true score and error score. Test scores may be expressed in the familiar equation:

$$\text{Observed Score} = \text{True Score} + \text{Error}$$

In this framework, the *observed score* is the test score that was actually obtained. The *true score* is the hypothetical amount of the designated trait specific to the examinee, a quantity that would be expected if the entire universe of relevant content were assessed or if the examinee were tested an infinite number of times without any confounding effects of such things as practice or fatigue. *Measurement error* is defined as the difference between true score and observed score. Error is uncorrelated with the true score and with other variables, and it is distributed normally and uniformly about the true score. Because its influence is random, the average measurement error across many testing occasions is expected to be zero.

Many of the key elements from contemporary psychometrics may be derived from this core assumption. For example, internal consistency reliability is a psychometric function of random measurement error, equal to the ratio of the true score variance to the observed score variance. By comparison, validity depends on the extent of nonrandom measurement error. Systematic sources of measurement error negatively influence validity, because error prevents measures from validly representing what they purport to assess. Issues of test fairness and bias are sometimes considered to constitute a special case of validity in which systematic sources of error across racial and ethnic groups constitute threats to validity generalization.

CTT places more emphasis on test score properties than on item parameters. According to Gulliksen (1950), the essential item statistics are the proportion of persons answering each item correctly (item difficulties, or *p*-values), the point-biserial correlation between item and total score multiplied by the item standard deviation (reliability index), and the point-biserial correlation between item and criterion score multiplied by the item standard deviation (validity index).

As a critic, Borsboom (2006a, 2006b; Borsboom, Mellenbergh, & Van Heerden, 2004) has argued that CTT has grave limitations in theory and model building through its misplaced emphasis on observed scores and true scores rather than the latent trait itself. Moreover, he has argued that it thereby creates a never-ending black hole need for continued accumulation of construct validity evidence (Borsboom 2006a, p. 431). At a more specific level, Hambleton, Swaminathan, and Rogers (1991) identified four limitations of CTT: (1) it has limited utility for constructing tests for dissimilar examinee populations (*sample dependence*); (2) it is not amenable for making comparisons of examinee performance on different tests purporting to measure the trait of interest (*test dependence*); (3) it operates under the assumption that equal measurement error exists for all examinees (*invariant reliability*); and (4) it provides no basis for predicting the likelihood of a given response of an examinee to a given test item, based on responses to other items. In general, with CTT, it is difficult to separate examinee characteristics from test characteristics. IRT addresses many of these limitations.

Item Response Theory

IRT may be traced to two separate lines of development. Its origins may be traced to the work of Danish mathematician Georg Rasch (1960), who developed a family of IRT models that separated person and item parameters. Rasch influenced the thinking of leading European and American psychometricians such as Gerhard Fischer and Benjamin Wright. A second line of development stemmed from research at the Educational Testing Service that culminated in Frederick Lord and Melvin Novick's (1968) classic textbook, including four chapters on IRT written by Allan Birnbaum. This book provided a unified statistical treatment of test theory and moved beyond Gulliksen's earlier CTT work.

IRT addresses the issue of how individual test items and observations map in a linear manner onto a targeted construct (termed the *latent trait,* with the amount of the trait denoted by θ). The frequency distribution of a total score, factor score, or other trait estimates is calculated on a standardized scale with a mean θ of 0 and a standard deviation (*SD*) of 1. An item response function (IRF; also known as an item characteristic curve, ICC) can then be created by plotting the proportion of people who have a score at each level of θ, so that the probability of a person's passing an item depends solely on the ability of that person and the properties of the item. The IRF curve yields several parameters, including item difficulty and item discrimination. Item *difficulty* is the location on the latent trait continuum corresponding to chance responding or, alternatively, the probability of responding accurately (or not) given a specified ability level. Item *discrimination* is the rate or slope at which the probability of success changes with trait level (i.e., the ability of the item to differentiate those with more of the trait from those with less). A third parameter denotes the probability of answering correctly by *guessing in low-ability respondents* (as with multiple-choice tests). A fourth parameter describes the probability of *carelessness in high-ability respondents* (i.e., those may answer an easy item incorrectly). IRT based on the one-parameter model (i.e., item difficulty) assumes equal discrimination for all items and negligible probability of guessing, and is generally referred to as the Rasch model. Two-parameter models (those that estimate both item difficulty and discrimination) and three-parameter models (those that estimate item difficulty, discrimination, and probability of guessing) may also be used. Only now are four-parameter models being considered of potential value, especially for their relevance in psychopathology (e.g., Loken & Rulison, 2010).

IRT posits several assumptions: (1) *unidimensionality and stability* of the latent trait, which is usually estimated from an aggregation of individual items; (2) *local independence* of items, meaning that the only influence on item responses is the latent trait and not adjacent (or any other) items; and (3) *item parameter invariance*—that is, item properties are a function of the item itself rather than the sample, test form, or interaction between item and respondent. Knowles and Condon (2000) argued that these assumptions may not always be made safely. While IRT offers technology that makes test development more efficient than CTT, its potential to lead to future advances in psychometrics is questionable. As Wainer (2010) asked rhetorically, "What more do we need to know about IRT to be able to use it well?" (p. 18).

SAMPLING AND NORMING

Under ideal circumstances, individual test results would be referenced to the performance of the entire collection of individuals (*target population*) for whom the test is intended. Statewide educational tests given to all students at specified grade levels have this potential, although they are often used only as criterion-referenced benchmarks of academic progress. Without government mandates, it is rarely feasible to measure performance of every member in a population. Accordingly, standardized tests are developed with the use of *sampling* procedures designed to provide an unbiased estimation of the score distribution and characteristics of a target population within a subset of individuals randomly selected from that population. Test results may then be interpreted with reference to sample characteristics, which are presumed to accurately estimate stable population parameters.

Appropriate Samples for Test Applications

When a test is intended to yield information about examinees' standing relative to peers of some kind, the chief objective of sampling should be to provide a reference group that is representative of the greater population for whom the test is intended. *Norm-referenced* test scores provide information about an examinee's standing relative to the distribution of test scores found in an appropriate peer comparison group. As a point of comparison, *criterion-referenced* tests yield scores that are interpreted relative to predetermined standards of performance, such as proficiency at a specific academic skill or activity of daily life.

If the test is intended to compare the performance of an individual to that of the general population, the sample may need to be stratified on demographic variables, typically those that account for substantial variation in test performance. Stratification divides the target population into smaller subpopulations, which can then be randomly sampled, provided that the population proportions in the strata are known (Kalton, 1983). Variables unrelated to the trait being assessed need not be included in the sampling plan. For example, a developmentally sensitive test needs to be stratified by age level, but a test that shows little variation in performance as a function of maturation may cover broad age ranges and not require age stratifications. The advantages of stratified sampling include greater ease of sampling and estimation, improved likelihood of representing important subpopulations in the normative sample, the option to conduct additional analyses on samples within strata (if independent from each other), and enhancement of sampling precision (e.g., Lehtonen & Pahkinen, 2004).

Variables frequently used for sample stratification include:

- Sex (female, male)
- Race (White, African American, Asian/Pacific Islander, Native American, Other)
- Ethnicity (Hispanic origin, non-Hispanic origin)
- Geographic region (Midwest, Northeast, South, West)
- Community setting (urban/suburban, rural)
- Parent educational attainment (less than high school degree, high school graduate or equivalent, some college or technical school, 4 or more years of college)

The most challenging of stratification variables is socioeconomic status (SES), particularly because it tends to be associated with cognitive test performance and is difficult to define operationally (e.g., Oakes & Rossi, 2003). Parent educational attainment is often used as an estimate of SES because it is readily available and objective and because parent education correlates moderately with family income. For children's measures, parent occupation and income are also sometimes combined as estimates of SES, although income information is generally difficult to obtain. Community estimates of SES add an additional level of sampling rigor, because the community in which an individual lives may be a greater factor in the child's everyday life experience than his or her parents' educational attainment. Similarly, the number of people residing in the home or whether one or two parents head the family are all factors that can influence a family's SES. For example, a family of three that has an annual income of $40,000 may have more economic viability than a family of six that earns the same income. Also, a college-educated single parent may earn less income than two lesser-educated cohabiting parents. The influences of SES on construct development clearly represent an area of further study, even as the relation of SES to cognitive and behavioral outcome proves complex (e.g., Turkheimer, Haley, Waldron, D'Onofrio, & Gottesman, 2003).

A classic example of an inappropriate normative reference sample is found with the original Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1943), which was normed on 724 Minnesota White adults who were, for the most part, relatives or visitors of patients in the University of Minnesota Hospitals. The original MMPI normative reference group was primarily composed of Minnesota farmers.

When test users intend to rank individuals relative to the special populations to which they belong, it may also be desirable to ensure that proportionate representation of those special populations are included in the normative sample (e.g., individuals who are intellectually disabled, conduct disordered, or learning disabled). Alternatively, it is not unusual to collect norms on special reference groups, such as individuals with a known diagnosis (e.g., autism spectrum disorders), when level of test performance is important in understanding the nature and severity of any impairments (e.g., specification of high-functioning autism versus lower-functioning autism). Millon, Davis, and Millon (1997) noted that tests normed on special populations may require the use of base rate scores rather than traditional standard scores, because assumptions of a normal distribution of scores often cannot be met within clinical populations.

Appropriate Sampling Methodology

One of the principal objectives of sampling is to ensure that each individual in the target population has an equal and independent chance of being selected. Sampling methodologies include both probability and nonprobability approaches, which have different strengths and weaknesses in terms of accuracy, cost, and feasibility (Levy & Lemeshow, 1999).

Probability sampling is a randomized approach that permits the use of statistical theory to estimate the properties of sample estimators. Probability sampling is generally too expensive for norming educational and psychological tests, but it offers the advantage of permitting the determination of the degree of sampling error, such as is frequently reported with the results of most public opinion polls. *Sampling error* may be defined as the difference between a sample statistic and its corresponding population parameter. When sampling error in psychological test norms is not reported, an important source of true score error that transcends measurement error alone will be neglected.

A probability sampling approach sometimes employed in psychological test norming is known as *multistage stratified random cluster sampling*; this approach uses a sampling strategy in which a large or dispersed population is divided into a large number of groups, with participants in the groups selected via random sampling. In two-stage cluster sampling, each group undergoes a second round of simple random sampling based on the expectation that each cluster closely resembles every other cluster. For example, a set of schools may constitute the first stage of sampling, with students randomly drawn from the schools in the second stage. Cluster sampling is more economical than random sampling, but incremental amounts of error may be introduced at each stage of sample selection. Moreover, cluster sampling commonly results in high standard errors when cases from a cluster are homogeneous (Levy & Lemeshow, 1999). Sampling error can be estimated with the cluster sampling approach, so long as the selection process at the various stages involves random sampling.

In general, sampling error tends to be largest when nonprobability-sampling approaches, such as convenience sampling or quota sampling, are employed *Convenience samples* involve the use of a self-selected sample that is easily accessible (e.g., volunteers, college subject pool participants, or examinees personally known to the examiner). *Quota samples* involve the selection by a coordinator of a predetermined number of cases with specific characteristics. The probability of acquiring an unrepresentative sample is high when using nonprobability procedures. The weakness of all nonprobability-sampling methods is that norms may not be applicable to the population being served, statistical theory cannot be used to estimate sampling precision, the likelihood of sampling bias is elevated, and accordingly sampling accuracy can be evaluated only subjectively (e.g., Kalton, 1983).

An example of best practice in sampling may be found in the approach used with the Achenbach System of Empirically Based Assessment (ASEBA; Achenbach & Rescorla, 2001), a family of scales that includes the Child Behavior Checklist (CBCL/6–18), a leading behavior rating scale for children and adolescents with behavior problems. A multistage national probability sample was collected in the process of updating the norms for these scales:

- One hundred primary sampling units were selected by an institute of survey research to be collectively representative of the U.S. population.
- Trained interviewers were assigned to households across the various sampling units to visit homes to determine the age and gender of residents eligible for participation (i.e., children and adolescents with no major physical or intellectual disability, with one English-speaking parent).
- Eligible residents were identified, and candidate participants were selected by stratified randomized procedures to match an overall target demographic specification, with no more than one candidate included from each household.
- Interviews with parents or youths were conducted to complete the rating scales.
- After receipt of completed scales, ASEBA staff telephoned respondents to verify that interviews actually had been conducted.

A completion rate of 93.0% was reported for eligible CBCL/6–18 participants, suggesting a low likelihood for sampling selection bias. Of the 2,029 children whose parents completed the CBCL/6–18, 276 (13.6%) were excluded after data collection based on parent reports of mental health, substance abuse, and special education services, yielding a final nonreferred normative sample of $N = 1{,}753$. The systematic and random sampling techniques used to norm the ASEBA behavior rating scales may be contrasted with the less randomized sampling techniques found with many other psychological tests and behavior rating scales.

Adequately Sized Normative Samples

How large should a normative sample be? If population parameters are to be estimated, effect sizes are to be calculated, or specific hypotheses are to be tested with null hypothesis significance testing, minimal sample sizes may be specified. The number of participants sampled at any given stratification level needs to be sufficiently large to provide acceptable sampling error with stable parameter estimations for the target populations. Depending on how the data are to be used, the *alpha level, effect size,* and *power* need to be specified in advance and can drive determination of minimal sample sizes necessary.

The minimum number of cases to be collected (or clusters to be sampled) also depends in part on the sampling procedure used; Levy and Lemeshow (1999) provided formulas for a variety of sampling procedures. Up to a point, the larger the sample the greater the reliability of sampling accuracy and the more precise the parameter estimate. Estimates that are biased will generally become less biased as sample size increases (e.g., Kelley & Rausch, 2006). Cattell (1986) noted that eventually diminishing returns can be expected when sample sizes are increased beyond a reasonable level. Julious (2005) recommended the use of a cost–benefit analyses, where the point at which increased sample size yields diminished effect in estimating relevant population parameters.

The smallest acceptable number of cases in a sampling plan may also be driven by the particular statistical analyses to be conducted. Hertzog (2008) recommended samples of $n = 25$ to 40 for pilot studies during instrument development, $n = 20$ to 25 for intervention efficacy pilot studies (capable of detecting large effect sizes), and $n = 30$ to 40 per group for pilot studies comparing groups. In contrast, Zieky (1993) recommended that a minimum of 500 examinees be distributed across the two groups compared in differential item function studies for group administered tests. For individually administered tests, differential item function analyses require substantial oversampling of minorities. With regard to exploratory factor analyses, Riese, Waller, and Comrey (2000) have reviewed the psychometric literature and concluded that most rules of thumb pertaining to minimum sample size are not useful. They suggested that when communalities are high and factors are well defined, sample sizes of 100 are often adequate, but when communalities are low, the number of factors is large, and the number of indicators per factor is small, even a sample size of 500 may be inadequate. As with statistical analyses in general, minimal acceptable sample sizes should be based on practical considerations, including such considerations as desired effect size, power, and alpha level.

As rules of thumb, group-administered tests undergoing standardization generally sample over 10,000 participants per age or grade level, whereas individually administered tests typically sample 100 to 200 participants per level (e.g., Robertson, 1992). In IRT, the minimum sample size is related to the choice of calibration model used. In an integrative review, Suen (1990) recommended that a minimum of 200 participants be examined for the one-parameter Rasch model, at least 500 examinees be examined for the two-parameter model, and at least 1,000 examinees be examined for the three-parameter model. Using the WISC-IV normative data set, Zhu and Chen (2011) reported that representative sample sizes as small as $N = 50$ per age cohort were capable of yielding comparable (or even improved) norms relative to a substantially larger sample, based on results with an inferential norming method (Wilkins & Rolfhus, 2004).

Confidence in the use of smaller normative samples may soon be enabled by advances in data resampling procedures (with replacement), such as the bootstrap, the jackknife, and permutation methods, that have been shown to provide stable estimates of statistical parameters without requiring assumptions as to normality or homogeneity of variance. In particular, the bootstrap technique has been utilized in two recent normative updates for the Cognitive Abilities Test (CogAT, Form 6; see Lohman & Lakin, 2009) and the Woodcock-Johnson III (WJ III; see McGrew, Dailey, & Schrank, 2007). Efron and Tibshirani (1993) described how the bootstrap might be used to construct a 95% confidence interval (CI) for the latent trait statistical parameter θ:

1. Draw 1,000 bootstrap samples with replacement from the original sample, each time calculating an estimate of θ.
2. Use the results to generate a (simulated) distribution of θ, sorting these estimates in ascending order.
3. Calculate the 2.5th percentile (i.e., the average of the 25th and 26th observations) and the 97.5th percentile (i.e., the average of the 975th and 976th observations) from the 1,000 simulated values.
4. The resulting values form the lower confidence limit and the upper confidence limit.

Mooney and Duval (1993) considered bootstrapped approximations of parameter estimates and CIs to be relatively high quality "when n reaches the range of 30–50, and when the sampling procedure is truly random" (p. 21).

Sampling Precision

As we have discussed, sampling error and bias is difficult to ascertain when probability sampling approaches are not used, and most educational and psychological tests do not employ true probability sampling. Given this limitation, there are few objective standards for the sampling precision of test norms. Angoff (1984) recommended as a rule of thumb that the maximum tolerable sampling error should be no more than 14% of the standard error of measurement. He declined, however, to provide further guidance in this area: “Beyond the general consideration that norms should be as precise as their intended use demands and the cost permits, there is very little else that can be said regarding minimum standards for norms reliability” (p. 79). For large-scale assessments normed through two-stage cluster sampling, a conventional recommendation has been that the magnitude of the 95% CI around a mean score should be less than 10% of the score’s SD (e.g., Foy & Joncas, 2004; Wu, 2010). CIs that are greater than 10% of the SD may indicate problematic deviations from random sampling, inadequate sample size, and insufficient power. Wu (2010) noted that large sampling error can easily account for spuriously large differences in group mean scores, such as those that might be expected between regions or over time as a product of instruction.

In the absence of formal estimates of sampling error, the accuracy of sampling strata may be most easily determined by comparing stratification breakdowns against those available for the target population. As the sample more closely matches population characteristics, the more representative is a test’s normative sample. As best practice, we recommend that test developers provide tables showing the composition of the standardization sample within and across all stratification criteria (e.g., Percentages of the Normative Sample according to combined variables, such as Age, by Race, or by Parent Education). This level of stringency and detail ensures that important demographic variables are distributed proportionately across other stratifying variables according to population proportions. The practice of reporting sampling accuracy for single-stratification variables “on the margins” (i.e., by one stratification variable at a time) tends to conceal lapses in sampling accuracy. For example, if sample proportions of low SES are concentrated in minority groups (instead of being proportionately distributed across majority and minority groups), then the precision of the sample has been compromised through the neglect of minority groups with high SES and majority groups with low SES. The more the sample deviates from population proportions on multiple stratifications, the greater the effect of sampling error.

Manipulation of the sample composition to generate norms is often accomplished through sample weighting (i.e., application of participant weights to obtain a distribution of scores that is exactly proportioned to the target population representations). Weighting is used more frequently with group-administered educational tests, because educational tests typically involve the collection of thousands of cases. Weighting is used less frequently with psychological tests, and its use with these smaller samples may significantly affect systematic sampling error because fewer cases are collected and therefore weighting may differentially affect proportions across different stratification criteria, improving one at the cost of another. Weighting is most likely to contribute to sampling error when a group has been inadequately represented with too few cases collected.

An illustration of problematic reporting of sample weighting may be found in the Wechsler Memory Scale (WMS-III; Wechsler, 1997). While this test’s technical manual reports a standardization sample of 1,250 examinees (Tulsky, Zhu, & Ledbetter, 1997), subsequent independent reports indicated that this was a “weighted” *N* and that 217 or 218 participants were exact duplicates of participants in the “unweighted” *N* of 1,032 (see Tulsky, Chiaravallotti, Palmer, & Chelune, 2003; also Frisby & Kim, 2008). This approach to weighting a normative sample is not clearly disclosed in test technical materials (see, e.g., *Standards for Educational and Psychological Testing,* 1999) and does not meet accepted weighting procedures (e.g., Rust & Johnson, 1992).

Recency of Sampling

How old can norms be and still remain accurate? Evidence from the last two decades suggests that norms from measures of cognitive ability are susceptible to becoming “soft” or “stale” (i.e., test consumers should use older norms with caution). Use of outdated normative samples introduces systematic error into the diagnostic process and may negatively influence decision making, such as denying services (for mentally handicapping conditions) to sizable numbers of children and adolescents who otherwise would have been identified as eligible to receive services (e.g., Reschly, Myers, & Hartel, 2002). Sample recency is an ethical concern for all psychologists who test or conduct assessments. The American Psychological Association’s (2002) Ethical Principles and Code of Conduct directs psychologists to avoid basing decisions or recommendations on results that stem from obsolete or outdated tests.

The problem of normative obsolescence has been most robustly demonstrated with intelligence tests. The term *Flynn effect* (Herrnstein & Murray, 1994) is used to describe a consistent pattern of population intelligence test score gains over time and across nations (Flynn, 1984, 1987, 1994, 1999). For intelligence tests, the rate of gain is about one-third of an IQ point per year (3 points per decade), which has been a roughly uniform finding over time and for all ages (Flynn, 1999). The Flynn effect appears to occur as early as infancy (Bayley, 1993; Campbell, Siegel, Parr, & Ramey, 1986) and continues through the full range of adulthood (Tulsky & Ledbetter, 2000). The effect implies that older test norms may yield inflated scores relative to current normative expectations. For example, the Wechsler Intelligence Scale for Children—III (WISC-III; Wechsler, 1991) currently yields higher Full Scale IQs than the fourth edition of the WISC (Wechsler, 2003) by about 2.5 IQ points.

How often should tests be revised? There is no empirical basis for making a global recommendation, but it seems reasonable to conduct normative updates, restandardizations, or revisions at time intervals corresponding to the time expected to produce 1 standard error of measurement (*SEM*) of change. For example, given the Flynn effect and an overall average WISC-IV Full Scale IQ *SEM* of 2.68, one could expect about 10 years to elapse before the test's norms would "soften" to the magnitude of 1 *SEM* . We note, however, that some evidence has suggested that the Flynn effect may have diminished or even reversed in recent years (e.g., Teasdale & Owen, 2005).

CALIBRATION AND DERIVATION OF REFERENCE NORMS

This section describes several psychometric characteristics of test construction as they relate to building individual scales and developing appropriate norm-referenced scores. *Calibration* refers to the analysis of properties of gradation in a measure, defined in part by properties of test items. *Norming* is the process of using scores obtained by an appropriate sample to build quantitative references that can be used effectively in the comparison and evaluation of individual performances relative to "typical" peer expectations.

Calibration

The process of item and scale calibration dates back to the earliest attempts to measure temperature. Early in the 17th century, there was no method to quantify heat and cold except through subjective judgment. Galileo and others experimented with devices that expanded air in glass as heat increased; use of liquid in glass to measure temperature was developed in the 1630s. Some two dozen temperature scales were available for use in Europe in the 17th century, and each scientist had his own scales with varying gradations and reference points. It was not until the early 18th century that more uniform scales were developed by Fahrenheit, Celsius, and de Réaumur.

The process of calibration has similarly evolved in psychological testing. In CTT, item difficulty is judged by the *p*-value, or the proportion of people in the sample that passes an item. During ability test development, items are typically ranked by *p*-value or the amount of the trait being measured. The use of regular, incremental increases in item difficulties provides a methodology for building scale gradations. Item difficulty properties in CTT are dependent on the population sampled, so that a sample with higher levels of the latent trait (e.g., older children on a set of vocabulary items) would show different item properties (e.g., higher *p*-values) than a sample with lower levels of the latent trait (e.g., younger children on the same set of vocabulary items).

In contrast, IRT includes both item properties and levels of the latent trait in analyses, permitting item calibration to be sample independent. The same item difficulty and discrimination values will be estimated regardless of trait distribution. This process permits item calibration to be sample free, according to Wright (1999), so that the scale transcends the group measured. Embretson (1999) has described one of the new rules of measurement: "Unbiased estimates of item properties may be obtained from unrepresentative samples" (p. 13).

IRT permits several item parameters to be estimated in the process of item calibration. Among the indices calculated in widely used Rasch model computer programs (e.g., Linacre & Wright, 1999) are item fit-to-model expectations, item difficulty calibrations, item-total correlations, and item standard error. The conformity of any item to expectations from the Rasch model may be determined by examining item fit. Items are said to have good fits with typical item characteristic curves when they show expected patterns near to and far from the latent trait level for which they are the best estimates. Measures of item difficulty adjusted for the influence of sample ability are typically expressed in logits, permitting approximation of equal difficulty intervals.

Item and Scale Gradients

The *item gradient of a test* refers to how steeply or gradually items are arranged by trait level and the resulting gaps that may ensue in standard scores. In order for a test to have adequate sensitivity to differing degrees of ability or any trait being measured, it must have adequate item density across the distribution of the latent trait. The larger the resulting standard score differences in relation to a change in a single raw score point, the less sensitive, discriminating, and effective a test is.

For example, on the Memory subtest of the Battelle Developmental Inventory (Newborg, Stock, Wnek, Guidubaldi, & Svinicki, 1984), a child who is 1 year 11 months old who earned a raw score of 7 would have performance ranked at the first percentile for age, while a raw score of 8 leaps to a percentile rank of 74. The steepness of this gradient in the distribution of scores suggests that this subtest is insensitive to even large gradations in ability at this age.

A similar problem is evident on the Motor Quality index of the Bayley Scales of Infant Development Behavior Rating Scale (BSID-II; Bayley, 1993). A 36-month-old child with a raw score rating of 39 obtains a percentile rank of 66. The same child obtaining a raw score of 40 is ranked at the 99th percentile.

As a recommended guideline, tests may be said to have adequate item gradients and item density when there are approximately three items per Rasch logit, or when passage of a single item results in a standard score change of less than one third *SD* (0.33 *SD*) (Bracken, 1987). Items that are not evenly distributed in terms of the latent trait may yield steeper change gradients that will decrease the sensitivity of the instrument to finer gradations in ability.

Floor and Ceiling Effects

Do tests have adequate breadth, bottom and top? Many tests yield their most valuable clinical inferences when scores are extreme (i.e., for scores that are very low or very high). Accordingly, tests used for clinical purposes need sufficient discriminating power in the extreme ends of the distributions.

The floor of a test represents the extent to which an individual can earn appropriately low standard scores. A floor is usually considered the lowest, nonzero raw score that may be earned for any instrument; zero raw scores have ambiguous meaning, because a wide array of construct-irrelevant explanations can account for zero scores. An intelligence test intended for use in the identification of individuals diagnosed with intellectual disabilities must, by definition, extend at least 2 SDs below normative expectations (IQ < 70). In order to serve individuals with severe to profound intellectual disability, test scores must extend even further to more than 4 SDs below the normative mean (IQ < 40). Tests without a sufficiently low floor may not be useful for decision making for more severe forms of cognitive impairment.

A similar situation arises for test ceiling effects. An intelligence test with a ceiling greater than 2 SDs above the mean (IQ > 130) can identify most candidates for intellectually gifted programs. To identify individuals as exceptionally gifted (i.e., IQ > 160), a test ceiling must extend more than 4 SDs above normative expectations. There are several unique psychometric challenges to extending norms to these heights, but recently the publisher of the leading school-age intelligence test, the WISC-IV, increased its highest global scores beyond 150–160 to extended standard scores as high as 210 (Zhu, Cayton, Weiss, & Gabel, 2008). The Stanford-Binet, Fifth Edition also offered an Extended IQ Score (EXIQ) that extends up to 225 (Roid, 2003). These advances may enable the identification of exceptionally gifted students at levels not previously possible.

As a rule of thumb, tests used for clinical decision making should have floors and ceilings that differentiate the extreme lowest and highest 2% of the population from the middlemost 96% (Bracken, 1987, 1988). Tests with inadequate floors or ceilings are inappropriate for assessing children with known or suspected intellectual disability, intellectual giftedness, severe psychopathology, or exceptional social and educational competencies.

Derivation of Norm-Referenced Scores

IRT yields several different kinds of interpretable scores (e.g., Woodcock, 1999), only some of which are norm-referenced standard scores. Because most test users are most familiar with the use of standard scores, we focus on the process of arriving at this type of score. Transformation of raw scores to standard scores involves a number of decisions based on psychometric science and more than a little art.

The first decision involves the nature of raw score transformations, based on theoretical considerations (*Is the trait being measured thought to be normally distributed?*) and examination of the cumulative frequency distributions of raw scores within and across age groups (e.g., Daniel, 2007). The objective of this transformation is to preserve the shape of the raw score frequency distribution, including mean, variance, kurtosis, and skewness. *Linear transformations* of raw scores are based solely on the mean and distribution of raw scores and are commonly used when distributions are not normal; linear transformation assumes that the distances between scale points reflect true differences in the degree of the measured trait present. *Area transformations* of raw score distributions convert the shape of the frequency distribution into a specified type of distribution. When the raw scores are normally distributed, they may be transformed to fit a normal curve, with corresponding percentile ranks assigned in a way so that the mean corresponds to the 50th percentile, −1 *SD* and +1 *SD* correspond to the 16th and 84th percentiles respectively, and so forth. When the frequency distribution is not normal, it is possible to select from varying types of nonnormal frequency curves (e.g., Johnson, 1949) as a basis for transformation of raw scores or to use polynomial curve-fitting equations.

Following raw score transformations is the process of smoothing the curves. Data smoothing typically occurs within and across groups to correct for minor irregularities, presumably those irregularities that result from sampling fluctuations and error. Quality checking also occurs to eliminate vertical reversals (such as those within an age group, from one raw score to the next) and horizontal reversals (such as those within a raw score series, from one age to the next). Smoothing and elimination of reversals serve to ensure that raw score to standard score transformations progress according to growth and maturation expectations for the trait being measured.

Beyond computing norms one age group at a time, *continuous norming* (Gorsuch, 1983b; Gorsuch & Zachary, 1985) began as a way of using test scores across a large number of overlapping age groups to generate polynomial regression equations that could accurately capture the developmental progression of test scores. Continuous norming enabled improved estimation of raw to standard score transformations and age-based percentile ranks while minimizing the effects of sampling and artifactual irregularities. This approach has evolved into *continuous parameter estimation methods* (Gorsuch, 2010; Roid, 2010) that permit computerized estimation of statistical parameters such as mean, SD, and skewness for different samples as a function of salient population characteristics (e.g., gender, education, ethnicity, and age), thereby providing a context for transforming test raw scores to standard scores. Continuous parameter estimation methods may also be used to compute and model reliabilities, standard errors of measurement, and various forms of validity as a function of other variables (Gorsuch, 2010; Roid, 2010).

TEST SCORE VALIDITY

Validity traditionally has been concerned with the *meaning* of test scores, or whether a test measures what it purports to measure (e.g., Cronbach & Meehl, 1955; Kelley, 1927). In an influential definition that sought to unify all forms of test score validity under the umbrella of *construct validity* and extend its reach to encompass the applied use of test results and their interpretations, Messick (1989a) defined *validity* as “an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment” (p. 13; emphasis in original). From this mainstream perspective, validity involves the inferences made from test scores and is not inherent to the test itself (e.g., Cronbach, 1971; Sireci, 2009; *Standards for educational and psychological testing,* 1999).

Yet Borsboom and his colleagues (Borsboom, Cramer, Kievit, Scholten, & Franić, 2009; Borsboom et al., 2004) argued that the traditional concept of validity was always indefensible, if just because it focused on test scores and their interpretations, whether they made sense in terms of psychological theories, and even on the justifiability of social actions based on test scores, rather than the measurement tools per se. Stripping excess requirements but covering less ground, they proposed a narrower formulation: that “validity is a property of measurement instruments [that] codes whether these instruments are sensitive to variation in a targeted attribute” (Borsboom et al., 2009, p. 135).

Lissitz (2009) provided an accessible compilation of controversies in contemporary perspectives on test score validity, from mainstream concepts of validity, to calls for radical change, and to *applications-oriented* forms of validity. In recent years, mainstream notions of test score validity have increasingly relied on the ambiguous concept of *construct validity,* which has come to represent something of a bottomless pit in terms of the ongoing accumulation of evidence for the validity of a test. Consumers of psychological test results expect the tests to have broad and diverse foundations and to be applied interpretatively in a manner supported by research. In a test-centered society, the narrow and more radical definition of validity espoused by Borsboom and his colleagues has a purist quality that appears inadequate, given the expectations of test consumers. The applications-oriented perspective takes the very functional approach that the optimal array of evidence necessary to support test validity varies according to the nature and applications of the test.

From a mainstream perspective, evidence of test score validity may take different forms, many of which are detailed in this chapter, but ultimately they are all concerned with construct validity (Guion, 1977; Messick, 1995a, 1995b). Construct validity involves appraisal of a body of evidence determining the degree to which test score inferences are accurate, adequate, and appropriate indicators of the examinee's standing on the trait or characteristic measured by the test. Excessive narrowness or broadness in the definition and measurement of the targeted construct can threaten construct validity. The problem of excessive narrowness, or *construct underrepresentation,* refers to the extent to which test scores fail to tap important facets of the construct being measured. The problem of excessive broadness, or *construct irrelevance,* refers to the extent to which test scores that are influenced by unintended factors, including irrelevant constructs and test procedural biases.

Construct validity can be supported with two broad classes of evidence: *internal* and *external* validation, which parallel the classes of threats to validity of research designs (Campbell & Stanley, 1963; Cook & Campbell, 1979). Internal evidence for validity includes information intrinsic to the measure itself, including content, examinee response processes, and substantive and structural validation. External evidence for test score validity may be drawn from research involving independent, criterion-related data. External evidence includes convergent, discriminant, criterion-related, and consequential validation. This internal–external dichotomy with its constituent elements represents a distillation of concepts described by Anastasi and Urbina (1997); Jackson (1971); Loevinger (1957); Messick (1995a, 1995b); Millon et al. (1997); and Slaney and Maraun (2008), among many others.

Internal Evidence of Validity

Internal sources of validity include the intrinsic characteristics of a test, especially its content, assessment methods, structure, and theoretical underpinnings. In this section, several sources of evidence internal to tests are described—including content validity, substantive validity, and structural validity.

Content Validity

Content validity is the degree to which elements of a test, ranging from items to instructions, are relevant to and representative of varying facets of the targeted construct (Haynes, Richard, & Kubany, 1995). Content validity is typically established through the use of expert judges who review test content, but other procedures may also be employed (Haynes et al., 1995; Vogt, King & King, 2004). Hopkins and Antes (1978) recommended that tests include a table of content specifications, in which the facets and dimensions of the construct are listed alongside the number and identity of items assessing each facet. More recently, Mislevy and his colleagues (e.g., Mislevy & Haertel, 2006) have proposed *evidence-centered design* (ECD), in which test developers use model-based reasoning and data-based warrants to formulate evidentiary arguments logically connecting test substance to meaningful claims about examinee performance and proficiency. Through this approach, for example, educational test items are written with the sole intent of eliciting explicitly defined forms of evidence to support inferences of interest, such as student mastery of a specific academic curriculum. According to Brennan (2010a), the validity claims of ECD, including the implied assertion that ECD-developed tests have validity built into the test a priori, need to be more rigorously substantiated.

Content differences across tests purporting to measure the same construct can explain why similar tests sometimes yield dissimilar results for the same examinee (Bracken, 1988). For example, the universe of mathematical skills includes varying types of numbers (e.g., whole numbers, decimals, fractions), number concepts (e.g., half, dozen, twice, more than), and basic operations (addition, subtraction, multiplication, division). The extent to which tests differentially sample content can account for differences between tests that purport to measure the same construct.

Tests should ideally include enough diverse content to adequately sample the breadth of construct-relevant domains, but content sampling should not be so diverse that scale coherence and uniformity is lost. Construct underrepresentation, stemming from use of narrow and homogeneous content sampling, tends to yield higher reliabilities than tests with heterogeneous item content, at the potential cost of generalizability and external validity. In contrast, tests with more heterogeneous content may show higher validity with the concomitant cost of scale reliability. Clinical inferences made from tests with excessively narrow breadth of content may be suspect, even when other indices of validity are satisfactory (Haynes et al., 1995).

Content validity is particularly valued for educational achievement testing, vocational testing, and some self-report measures of personality and psychopathology, because the congruence of test content with test interpretation is quite linear. For state educational assessments, for example, Crocker (2003) observed that the match between test content and curricular requirements buttresses the legal defensibility of standardized tests: "When scores are used for educational accountability, the 'load-bearing wall' of that [validity] argument is surely content representativeness" (p. 7).

In the field of personality assessment, the development of the Minnesota Multiphasic Personality Inventory, Second Edition (MMPI-2) Restructured Clinical Scales and the subsequent publication of the MMPI-2 Restructured Form (MMPI-2-RF; Ben-Porath & Tellegen, 2008; Tellegen & Ben-Porath, 2008) has marked a shift toward improved content validity in this most widely used personality test, because empirically keyed "subtle" items and items saturated with a primary "demoralization" factor have been removed from the clinical scales, eliminating some 40% of MMPI-2 items to generate a new form, the MMPI-2-RF, in which item content is easily mapped onto the clinical dimension being measured. This was not always true for previous editions of the MMPI, which included items with unclear content relevance for personality and psychopathology:

- I used to like drop-the-handkerchief.
- I liked *Alice's Adventures in Wonderland* by Lewis Carroll.

In a special journal issue dedicated to a possible paradigm shift away from MMPI and MMPI-2 empirically derived item traditions, Weed (2006) outlined the strengths, limitations, and controversies associated with the new MMPI-2-RF.

Examinee Response Processes

An examination of *how* individuals solve problems or answer questions is potentially important in establishing that a test measures what it purports to measure. The Standards for Educational and Psychological Testing (1999) stated that evidence based on response processes concerns "the fit between the construct and the detailed nature of performance or response actually engaged in by examinees" (p. 12). For example, a test measuring mathematical proficiency should not be difficult because its vocabulary is not understood; alternatively, a pictorial measure should not appear difficult merely because the picture is confusing. While item-level statistical analyses will normally identify test items that do not perform adequately from a psychometric perspective, new qualitative methods are being developed to identify the mental processes by which examinees understand and respond to test items.

The most compelling of these methods are Ericsson and Simon's (1980) "thinking aloud" procedures, which involve focusing on a challenging task while concurrently giving verbal expression to thoughts entering attention. A meta-analysis has shown that the think-aloud method shows no evidence of reactivity (or other influence on the accuracy of performance) (Fox, Ericsson, & Best, 2011). A representative think-aloud procedure consists of this directive given to a test taker:

> I would like you to start reading the questions aloud and tell me what you are thinking as you read the questions. After you have read the question, interpret the question in your own words. Think aloud and tell me what you are doing. What is the question asking you to do? What did you have to do to answer the question? How did you come up with your solution? Tell me everything you are thinking while you are doing the question. (Ercikan et al., 2011, p. 27)

For educational achievement testing, the think-aloud responses are recorded and scored according to four themes: understanding of the item, difficulty of the item, aspects of the item that are helpful in arriving at a solution, and aspects of the item that are confusing and difficult to understand (Ercikan, Arim, Law, Domene, Gagnon, & Lacroix, 2011).

Other methods to study examinee response processes include interviews of examinees, observation of test session behaviors, examination of item reaction times, eye movement tracking, and even simultaneous functional brain mapping methodologies. For our narrow goal of improving test score validity, think aloud protocols represent an economical way to ensure that test items are eliciting responses that tap the targeted construct and not construct-irrelevant responses.

Substantive Validity

The formulation of test items and procedures based on and consistent with a theory has been termed *substantive validity* (Loevinger, 1957). The presence of an underlying theory enhances a test's construct validity by providing scaffolding between content and constructs, which logically explains relations between elements, predicts undetermined parameters, and explains findings that would be anomalous within another theory (e.g., Kuhn, 1970). As Crocker and Algina (1986) suggested, "[P]sychological measurement, even though it is based on observable responses, would have little meaning or usefulness unless it could be interpreted in light of the underlying theoretical construct" (p. 7).

Many major psychological tests remain psychometrically rigorous but impoverished in terms of theoretical underpinnings. For example, conspicuously little theory is associated with most widely used measures of intelligence (e.g., the Wechsler scales), behavior problems (e.g., the Child Behavior Checklist), neuropsychological functioning (e.g., the Halstead-Reitan Neuropsychology Battery), and personality and psychopathology (the MMPI-2). It may well be that there are post hoc benefits to tests developed without theories; as observed by Nunnally and Bernstein (1994), “Virtually every measure that became popular led to new unanticipated theories” (p. 107). Moreover, tests with well-articulated theories may be easier to discredit than tests without accompanying theory—because falsifying the theory will undermine the substantive validity of the test.

Personality assessment has taken a leading role in theory-based test development, while only now, with the rise of the Cattell-Horn-Carroll framework for understanding human abilities, is cognitive-intellectual assessment increasingly relying of theory. Describing best practices for the measurement of personality some three decades ago, Loevinger (1972) commented, “Theory has always been the mark of a mature science. The time is overdue for psychology, in general, and personality measurement, in particular, to come of age” (p. 56).

Structural Validity

Structural validity relies mainly on factor-analytic techniques to identify a test’s underlying dimensions and the variance associated with each dimension. Also called *factorial validity* (Guilford, 1950), this form of validity may utilize other methodologies, such as multidimensional scaling, to help researchers understand a test’s structure. Structural validity evidence is generally internal to the test, based on the analysis of constituent subtests or scoring indices. Structural validation approaches may also combine two or more instruments in cross-battery factor analyses to explore evidence of convergent validity.

The two leading factor analytic methodologies used to establish structural validity are exploratory and confirmatory factor analyses. Exploratory factor analyses (EFAs) allow for empirical derivation of the structure of an instrument, often without a priori expectations, and are best interpreted according to the “psychological meaningfulness” of the dimensions or factors that emerge (e.g., Gorsuch, 1983a). Confirmatory factor analyses (CFAs) help researchers evaluate the congruence of the test data with a specified model and measure the relative fit of competing models. Confirmatory analyses explore the extent to which the proposed factor structure of a test explains its underlying dimensions as compared to alternative theoretical explanations. Thompson (2004) asserted, “Both EFA and CFA remain useful today, and our selection between the two classes of factor analysis generally depends on whether we have specific theory regarding data structure” (p. 6).

As a recommended guideline, the underlying factor structure of a test should be congruent with its composite indices (e.g., Floyd & Widaman, 1995), and the interpretive structure of a test should be the best-fitting structural model available. For example, the transformation of the Wechsler intelligence scales from a historically dichotomous interpretive structure (i.e., verbal and performance IQ, subordinate to the Full Scale IQ) to a four-factor interpretive structure (i.e., the verbal comprehension index, perceptual organization index, working memory index, and processing speed index, each contributing to the superordinate Full Scale IQ) reflects the relative importance of factor-analytic studies in driving test design and structure (Wechsler, 2003, 2008).

In the areas of personality and psychopathology assessment, leading instruments have long been plagued by inadequate structural validity. The MMPI and its restandardization, the MMPI-2, have received highly critical reviews as being “suboptimal from the perspective of modern psychometric standards” (Helmes & Reddon, 1993, p. 453), particularly for the mismatch between their psychometric and interpretive structure (e.g., Horn, Wanberg, & Appel, 1973). Some improvement is reported in the factorial support for the MMPI-2-RF Restructured Clinical scales, which extracted items tapping demoralization content (that saturated the clinical scales) to an independent scale, thereby leaving the restructured clinical scales more unidimensional with higher reliability (Hoelzle & Meyer, 2008; Tellegen, Ben-Porath, McNulty, Arbisi, Graham, & Kaemmer, 2003).

A different problem with structural validity has been observed with the widely recognized “Big Five” five-factor model of normal range personality, which has been repeatedly supported with exploratory factor analyses and disconfirmed with confirmatory factor analyses (e.g., Gignac, Bates, & Jang, 2007; Vassend & Skrondal, 2011). The five-factor model represents personality trait structure in terms of five orthogonal factors—neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness—and is most commonly assessed with the NEO Personality Inventory (NEO-PI-3; Costa & McCrae, 2010). In response to consistent findings of poor model fit via CFA, McCrae, Zonderman, Costa, Bond, and Paunonen (1996) have asserted that confirmatory factor analysis is systematically flawed, capable of showing poor fits for reliable structures, and they warn of “the dangers in an uncritical adoption and simplistic application of CFA techniques” (p. 563).

The cases of the MMPI-2 and the NEO-PI-3 suggest that a reasonable balance needs to be struck between theoretical underpinnings and structural validation; that is, if factor-analytic techniques do not consistently support a test's underpinnings, further research is needed to determine whether that is due to limitations of the theory, the factor-analytic methods, the nature of the test, or a combination of these factors. Carroll (1983), whose factor-analytic work has been influential in contemporary cognitive assessment, cautioned against overreliance on factor analysis as principal evidence of validity, encouraging use of additional sources of validity evidence that move "beyond factor analysis" (p. 26).

External Evidence of Validity

Evidence of test score validity also includes the extent to which the test results predict meaningful and generalizable behaviors independent of actual test performance. Test results need to be validated for any intended application or decision-making process in which they play a part. This section describes external classes of evidence for test construct validity, including convergent, discriminant, criterion-related, and consequential validity, as well as specialized forms of validity within these categories.

Convergent and Discriminant Validity

In a classic 1959 article, Campbell and Fiske described a multitrait-multimethod methodology for investigating construct validity. In brief, they suggested that a measure is jointly defined by its methods of gathering data (e.g., self-report or parent report) and its trait-related content (e.g., anxiety or depression). They noted that test scores should be related to (i.e., strongly correlated with) other measures of the same psychological construct (*convergent* evidence of validity) and comparatively unrelated to (i.e., weakly correlated with) measures of different psychological constructs (*discriminant* evidence of validity). The multitrait-multimethod matrix allows for the comparison of the relative strength of association between two measures of the same trait using different methods (monotrait-heteromethod correlations), two measures with a common method but tapping different traits (heterotrait-monomethod correlations), and two measures tapping different traits using different methods (heterotrait-heteromethod correlations), all of which are expected to yield lower values than internal consistency reliability statistics using the same method to tap the same trait.

The multitrait-multimethod matrix offers several advantages, such as the identification of problematic method variance. *Method variance* is a measurement artifact that threatens validity by producing spuriously high correlations between similar assessment methods of different traits. For example, high correlations between digit span, letter span, phoneme span, and word span procedures might be interpreted as stemming from the immediate memory span recall method common to all the procedures rather than any specific abilities being assessed. Method effects may be assessed by comparing the correlations of different traits measured with the same method (i.e., monomethod correlations) and the correlations among different traits across methods (i.e., heteromethod correlations). Method variance is said to be present if the heterotrait-monomethod correlations greatly exceed the heterotrait-heteromethod correlations in magnitude, assuming that convergent validity has been demonstrated.

Fiske and Campbell (1992) subsequently recognized shortcomings in their methodology: "We have yet to see a really good matrix: one that is based on fairly similar concepts and plausibly independent methods and shows high convergent and discriminant validation by all standards" (p. 394). At the same time, the methodology has provided a useful framework for establishing evidence of validity.

Criterion-Related Validity

How well do test scores predict performance on independent criterion measures and differentiate criterion groups? The relationship of test scores to relevant external criteria constitutes evidence of *criterion-related validity,* which may take several different forms. Evidence of validity may include criterion scores that are obtained at about the same time (*concurrent* evidence of validity) or criterion scores that are obtained at some future date (*predictive* evidence of validity). External criteria may also include functional, real-life variables (*ecological* validity), diagnostic or placement indices (*diagnostic* validity), and intervention-related approaches (*instructional* and *treatment* validity).

The emphasis on understanding the functional implications of test findings describes *ecological validity* (Neisser, 1978). Banaji and Crowder (1989) suggested, "If research is scientifically sound it is better to use ecologically lifelike rather than contrived methods" (p. 1188). In essence, ecological validation efforts relate test performance to various aspects of person–environment functioning in everyday life, including identification of both competencies and deficits in social and educational adjustment. Test developers should show the ecological relevance of the constructs a test purports to measure as well as the utility of the test for predicting everyday functional limitations for remediation. In contrast, tests based on laboratory-like procedures with little or no discernible relevance to real life may be said to have little ecological validity.

The capacity of a measure to produce relevant applied group differences has been termed *diagnostic validity* (e.g., Ittenbach, Esters, & Wainer, 1997). When tests are intended for diagnostic or placement decisions, diagnostic validity refers to their utility in differentiating the groups of concern. The process of arriving at diagnostic validity may be informed by decision theory, which involves calculations of decision-making accuracy in comparison to the base rate occurrence of an event or diagnosis in a given population. Decision theory has been applied to psychological tests (Cronbach & Gleser, 1965) and other high-stakes diagnostic tests (Swets, 1992) and is useful for identifying the extent to which tests improve clinical or educational decision making.

Contrasted groups is a common methodology to demonstrate diagnostic validity. In this methodology, test performance of two samples that are known to be different on the criterion of interest is compared. For example, a test intended to tap behavioral correlates of anxiety should show differences between groups of "normal" individuals and individuals diagnosed with anxiety disorders. A test intended for differential diagnostic utility should be effective in differentiating individuals with anxiety disorders from diagnoses that appear behaviorally similar. Decision-making classification accuracy may be determined by developing cutoff scores or rules to differentiate the groups, so long as the rules show adequate sensitivity, specificity, positive predictive power, and negative predictive power, as defined next.

- *Sensitivity*: The proportion of cases in which a clinical condition is detected when it is in fact present (true positive)
- *Specificity*: The proportion of cases for which a diagnosis is rejected when rejection is in fact warranted (true negative)
- *Positive predictive power*: The probability of having the diagnosis given that the score exceeds the cutoff score
- *Negative predictive power*: The probability of not having the diagnosis given that the score does not exceed the cutoff score

All of these indices of diagnostic accuracy are dependent on the prevalence of the disorder and the prevalence of the score on either side of the cut point.

Findings pertaining to decision making should be interpreted conservatively and cross-validated on independent samples because (a) classification decisions should in practice be based on the results of multiple sources of information rather than test results from a single measure, and (b) the consequences of a classification decision should be considered in evaluating the impact of classification accuracy. A false negative classification, meaning a child is incorrectly classified as not needing special education services, could mean the denial of needed services to a student. Alternatively, a false positive classification, in which a typical child is recommended for special services, could result in a child being labeled unfairly.

Bayesian methods to calculate evidential probabilities hold considerable promise in enhancing applied decision making, by permitting prior probabilities to be specified and then updated with relevant data such as test results. For example, the Bayesian *nomogram* represents a simple and practical strategy that is empirically derived, flexible, and easy to use as an aid to clinical decision making; it enables base rate information and test findings to be integrated to arrive at the probability for any likely outcome (e.g., Bianchi, Alexander, & Cash, 2009; Jenkins, Youngstrom, Washburn, & Youngstrom, 2011).

Treatment validity refers to the value of an assessment in selecting and implementing interventions and treatments that will benefit the examinee. "Assessment data are said to be *treatment valid,*" commented Barrios (1988), "if they expedite the orderly course of treatment or enhance the outcome of treatment" (p. 34). Other terms used to describe treatment validity are *treatment utility* (Hayes, Nelson, & Jarrett, 1987) and *rehabilitation-referenced assessment* (Heinrichs, 1990).

Whether the stated purpose of clinical assessment is description, diagnosis, intervention, prediction, tracking, or simply understanding, its ultimate raison d'être is to select and implement services in the best interests of the examinee, that is to guide treatment. In 1957, Cronbach described a rationale for linking assessment to treatment: "For any potential problem, there is some best group of treatments to use and best allocation of persons to treatments" (p. 680).

The origins of treatment validity may be traced to the concept of aptitude by treatment interactions (ATIs) originally proposed by Cronbach (1957), who initiated decades of research seeking to specify relationships between the traits measured by tests and the intervention methodology used to produce change. In clinical practice, promising efforts to match client characteristics and clinical dimensions to preferred therapist characteristics and treatment approaches have been made (e.g., Beutler & Clarkin, 1990; Beutler & Harwood, 2000; Lazarus, 1973; Maruish, 1994), but progress has been constrained in part by difficulty in arriving at consensus for empirically supported treatments (e.g., Beutler, 1998). In psychoeducational settings, tests results have been shown to have limited utility in predicting differential responses to varied forms of instruction (e.g., Reschly, 1997).

Turning the model that test results can predict effective treatment upside-down, recent federal mandates in the reauthorized Individuals with Disabilities Education Act (IDEA) have led to the practice of identifying learning disorders by Response to Intervention (RTI). RTI addresses the educational needs of at-risk students by delivering a series of instructional interventions accompanied by frequent progress measurements; students who do not benefit are considered in need of special education and are referred for further assessment and intervention. More than anything else, in RTI, it is the inadequate response to the initial treatment (i.e., evidence-based forms of instruction) that becomes diagnostic of a learning problem and potentially qualifies a student for special education (e.g., Vaughn & Fuchs, 2003).

Consequential Validity

The most recently proposed source of evidence for test score validity is concerned with both the intended and the unintended effects of test usage on individuals and groups. Messick (1989a, 1989b, 1995b) argued that test developers must understand the social values intrinsic to the purposes and application of psychological tests, especially those that may act as a trigger for social and educational actions. In this context, *consequential validity* refers to the appraisal of value implications and the social and legal impact of score interpretation as a basis for action and labeling as well as the actual and potential consequences of test use (Messick, 1989a, 1989b; Reckase, 1998).

Legal rulings and legislative actions have played a substantial role in establishing consequential validity, which was addressed in the landmark legal case of *Larry P. v. Riles* (343 F. Supp. 306, 1972; 495 F. Supp. 926, 1979), in which the court wrote that the consequences of test usage must be aligned with valid interpretation about what the test purports to measure. More recently, the text of the federal No Child Left Behind Act of 2001 (NCLB, 2002) legislation included a validity clause stating that assessments must "be used for purposes for which such assessments are valid and reliable, and be consistent with relevant, nationally recognized professional and technical standards" [20 U.S.C. § 6311(b)(3)(C)(iii)(2002)]. For large-scale, high-stakes educational testing, this clause has been interpreted as taking on meanings associated with social and racial equity:

> Couched in this statutory framework, the exact meaning of the validity clause becomes extremely important. One possible definition of the validity clause would ensure only that NCLB tests have certain statistical relationships with behaviors that students exhibit in a non-test environment. Other possible definitions of the validity clause would control the quality of testing practices to a much greater extent. For example, some meanings of the validity clause would mandate consideration of whether NCLB testing practices disproportionately affect students on the basis of race. Other notions of the validity clause would go even further by considering how well NCLB accountability measures achieve their statutory goal of boosting academic achievement. (Superfine, 2004, p. 482)

The immediate effect of the validity clause appears to have been to codify requirements for proportionate classification of racial and ethnical minorities in special education, with applications to grade retention and promotion policies and a wide array of other educational practices. Linn (1998) suggested that when governmental bodies establish policies that drive test development and implementation, the responsibility for the consequences of test usage must also be borne by the policy makers—and this form of validity extends far beyond the ken of test developers.

Consequential validity represents an expansion of traditional conceptualizations of test score validity. Lees-Haley (1996) urged caution about consequential validity, noting its potential for encouraging the encroachment of politics into science. The *Standards for Educational and Psychological Testing* (1999) recognized but carefully circumscribed consequential validity:

> Evidence about consequences may be directly relevant to validity when it can be traced to a source of invalidity such as construct underrepresentation or construct-irrelevant components. Evidence about consequences that cannot be so traced—that in fact reflects valid differences in performance—is crucial in informing policy decisions but falls outside the technical purview of validity. (p. 16)

Evidence of consequential validity may be collected by test developers during a period starting early in test development and extending through the life of the test

(Reckase, 1998). For educational tests, surveys and focus groups have been described as two methodologies to examine consequential aspects of validity (Chudowsky & Behuniak, 1998; Pomplun, 1997). The extent to which test results yield statistical evidence of disparate or discriminatory results on protected groups may also constitute compelling evidence of test score consequential (in)validity with legal implications (e.g., Superfine, 2004). As the social consequences of test use and interpretation are ascertained, the development and determinants of the consequences need to be explored. A measure with unintended negative side effects calls for examination of alternative measures and assessment counterproposals. Consequential validity is especially relevant to issues of bias, fairness, and distributive justice.

After a comprehensive survey of validity research published or presented in the past decade, Cizek, Bowen, and Church (2010) reported that consequential validity research was "essentially nonexistent in the professional literature" (p. 732), leading them to call it "a flaw in modern validity theory" (p. 739). They hypothesized that it is not possible to include consequences of test usage as a logical part of validation, in part because of the difficulty of synthesizing and integrating consequential value judgments with more traditional psychometric data per se. Moreover, they recommended differentiating the validation of score inferences from justifications for test use.

Validity Generalization

The accumulation of external evidence of test validity becomes most important when test results are generalized across contexts, situations, and populations and when the consequences of testing reach beyond the test's original intent. According to Messick (1995b), "The issue of generalizability of score inferences across tasks and contexts goes to the very heart of score meaning. Indeed, setting the boundaries of score meaning is precisely what generalizability evidence is meant to address" (p. 745).

Hunter and Schmidt (1990; Hunter, Schmidt, & Jackson, 1982; Schmidt & Hunter, 1977) developed a methodology of *validity generalization,* a form of meta-analysis, that analyzes the extent to which variation in test validity across studies is due to sampling error or other sources of error such as imperfect reliability, imperfect construct validity, range restriction, or artificial dichotomization. Once incongruent or conflictual findings across studies can be explained in terms of sources of error, meta-analysis enables theory to be tested, generalized, and quantitatively extended.

TEST SCORE RELIABILITY

If measurement is to be trusted, then it must be *reliable*. It must be consistent, accurate, and uniform across testing occasions, across time, across observers, and across samples—at least to the extent that the trait or construct being measured is stable. In psychometric terms, *reliability* refers to the extent to which measurement results are precise and accurate, reproducible, and free from random and unexplained error. Reliability has been described as "fundamental to all of psychology" (Li, Rosenthal, & Rubin, 1996, p. 98), and its study dates back nearly a century (Brown, 1910; Spearman, 1910). Reliability is the ratio of true score variance to observed score variance or, alternatively, the squared correlation between true and observed scores (e.g., Lord & Novick, 1968). Although traditional statistics such as Cronbach's coefficient alpha remain preeminent in published reliability research (Hogan, Benjamin, & Brezinski, 2000), somewhat newer important concepts in reliability include generalizability theory (albeit several decades "new"; Cronbach, Gleser, Nanda, & Rajaratnam, 1972) and reliability generalization (Vacha-Haase, 1998), both of which have important implications for how reliability is reported (e.g., Fan & Thompson, 2001). In this section, reliability is described according to CTT and IRT. Guidelines are provided for the objective evaluation of reliability.

The idea that reliability is a fixed property of a test or scale has been described as the primary myth about reliability still ubiquitous in test manuals (e.g., Streiner, 2003). As articulated by Wilkinson and the APA Task Force on Statistical Inference (1999), "Reliability is a property of the scores on a test for a particular population of examinees" (p. 596). More specifically, reliability is dependent on total score variance, so factors such as sample composition and test score variability will affect score reliability. For a fixed error variance, reliability will generally be large for a heterogeneous sample with large true score variance but small for a more homogeneous sample with small true score variance (e.g., Meyer, 2010).

Because any selected research sample may have test score reliabilities that differ significantly from the score reliabilities reported for normative standardization samples in test manuals, psychometric authorities have recommended that reliability indexes be calculated anew and reported as new samples are collected for research (Vacha-Haase, Kogan, & Thompson, 2000). Moreover, it is helpful for normative samples to report not only score reliabilities for each age cohort and score but also for any sample subgroups that may exhibit different levels of heterogeneity.

Generalizability Theory

While CTT decomposes observed score variance into true score variance and undifferentiated random error variance, an extension of CTT termed *generalizability theory* (Cronbach et al., 1972) includes a family of statistical procedures that estimates and partitions multiple sources of measurement error variance (facets) and their interactions. Generalizability theory posits that a response score is defined by the specific conditions under which it is produced, such as scorers, methods, settings, and times, and employs analysis of variance (ANOVA) methods to untangle the error associated with each of these conditions (e.g., Brennan, 2010b; Cone, 1978). Generalizability theory provides two reliability indexes: a *generalizability coefficient* (analogous to a reliability coefficient in CTT, estimating relative reliability for a wide range of scores), and a *dependability coefficient* (an estimate of absolute reliability in making criterion-referenced decisions, such as the reliability of passing and failing an academic proficiency exam with a cut score). Thompson (2003) noted that generalizability theory has the advantages of simultaneously enabling quantification of (a) multiple sources of measurement error variance, (b) interactions that create additional sources of measurement error variance, and (c) reliability for different types of decisions (i.e., relative or absolute). In spite of the powerful techniques found in generalizability theory, its use appears conspicuously absent from mental measures created in recent years (Hogan et al., 2000).

Internal Consistency

Determination of a test's internal consistency addresses the degree of uniformity and coherence among its constituent parts. Tests that are more uniform and unidimensional tend to be more reliable. As a measure of internal consistency, the reliability coefficient is the square of the correlation between obtained test scores and true scores; it will be high if there is relatively little error but low with a large amount of error. In CTT, reliability is based on the assumption that measurement error is distributed normally and equally for all score levels. By contrast, IRT posits that reliability differs between persons with different response patterns and levels of ability but generalizes across populations (Embretson & Hershberger, 1999).

Several statistics typically are used to calculate internal consistency. The split-half method of estimating reliability effectively splits test items in half (e.g., odd items and even items) and correlates the score from each half of the test with the score from the other half. This technique reduces the number of items in the test, thereby reducing the magnitude of the reliability. Use of the Spearman-Brown prophecy formula permits extrapolation from the obtained reliability coefficient to original length of the test, typically raising the reliability of the test. By far the most common statistical index of internal consistency is Cronbach's alpha, which provides a lower-bound estimate of test score reliability equivalent to the average split-half consistency coefficient for all possible divisions of the test into halves (Hogan et al., 2000).

Several recent studies serve to elucidate the limitations of Cronbach's alpha, specifically that it is strongly affected by scale length, that a high score does not ensure scale unidimensionality, and that excessively high scores (>.90) on subtests or scales are potentially risky. The effects of scale length on alpha have been long known, but Cortina (1993) demonstrated that when item incorrelations are held constant, increasing the length of a scale will substantially (and spuriously) raise its coefficient alpha—even when scales consist of two or three uncorrelated subscales. Cortina concluded: "If a scale has more than 14 items, then it will have an α of .70 or better even if it consists of two orthogonal dimensions with modest (i.e., .30) item intercorrelations. If the dimensions are correlated with each other, as they usually are, then α is even greater" (p. 102).

Accordingly, alpha has limited value as a measure of scale unidimensionality or homogeneity (Cortina, 1993; Streiner, 2003). As alpha rises above .90, it becomes possible that its capacity to estimate high internal consistency may instead signal item redundancy (essentially the same content expressed with different verbiage), leading Streiner to caution against use of single scales and tests with an alpha great than .90. Clark and Watson (1995) concurred in principle, observing: "Maximizing internal consistency almost invariably produces a scale that is quite narrow in content; if the scale is narrower than the target construct, its validity is compromised" (pp. 316–317). Nunnally and Bernstein (1994, p. 265) stated more directly: "Never switch to a less valid measure simply because it is more reliable." Conversely, highly homogeneous item sets may evidence high reliability as a function of limited content or construct sampling. Table 3.1 provides practical guidelines for evaluating test reliability coefficients, with higher coefficients needed when high-stakes individual student decisions are to be made.

TABLE 3.1 Guidelines for Acceptable Internal Consistency Reliability Coefficients

Test Methodology	Purpose of Assessment	Median Reliability Coefficient
Group Assessment	Programmatic decision making	.60 or greater
Individual Assessment	Screening	.80 or greater
	Diagnosis, intervention, placement, or selection	.90 or greater

Local Reliability and Conditional Standard Error

Internal consistency indexes of reliability provide a single average estimate of measurement precision across the full range of test scores. This approach assumes that measurement error variance is similar for all scores, an assumption that is generally false (e.g., Dimitrov, 2002). In contrast, *local reliability* refers to measurement precision at specified trait levels or ranges of scores. *Conditional error* refers to the measurement variance at a particular level of the latent trait, and its square root is a conditional standard error. Whereas CTT posits that the standard error of measurement is constant and applies to all scores in a particular population, IRT posits that the standard error of measurement varies according to the test scores obtained by the examinee but generalizes across populations (Embretson & Hershberger, 1999). In Rasch scaling, Wright (2001) observed, "once the test items are calibrated, the standard error corresponding to every possible raw score can be estimated without further data collection" (p. 786). Accordingly, reliability may be determined locally for any location in the score distribution (and level of latent trait) through IRT.

As an illustration of the use of CTT in the determination of local reliability, the Universal Nonverbal Intelligence Test (UNIT; Bracken & McCallum, 1998) presents local reliabilities from a CTT orientation. Based on the rationale that a common cut score for classification of individuals as mentally retarded is an FSIQ equal to 70, the reliability of test scores surrounding that decision point was calculated. Specifically, coefficient alpha reliabilities were calculated for FSIQs from −1.33 and −2.66 *SD*s below the normative mean. Reliabilities were corrected for restriction in range, and results showed that composite IQ reliabilities exceeded the .90 suggested criterion. That is, the UNIT is sufficiently precise at this ability range to reliably identify individual performance near to a common cut point for classification as intellectually disabled.

Two recent investigations have provided evidence supporting the importance of conditional error variance. Hopwood and Richards (2005) reported an increased frequency of scoring errors for high-ability samples on the Wechsler Adult Intelligence Scale (WAIS-III; Wechsler, 1997). In a follow-up investigation, Erdodi, Richard, and Hopwood (2009) reported evidence of significantly greater scoring errors in high- and low-ability samples on the Wechsler Intelligence Scale for Children (WISC-IV; Wechsler, 2003). These findings suggest that local reliabilities may be more appropriate within specified ability ranges than the single reliability estimates more conventionally used.

IRT permits the determination of conditional standard error at every level of performance on a test. Several measures, such as the Differential Ability Scales (Elliott, 1990) and the Scales of Independent Behavior (SIB-R; Bruininks, Woodcock, Weatherman, & Hill, 1996), report local standard errors or local reliabilities for every test score. This methodology not only determines whether a test is more accurate for some members of a group (e.g., high-functioning individuals) than for others (Daniel, 1999) but also promises that many other indexes derived from reliability indexes (e.g., index discrepancy scores) may eventually be tailored to an examinee's actual performance. Several IRT-based methodologies are available for estimating local scale reliabilities using conditional *SEM*s (Andrich, 1988; Daniel, 1999; Kolen, Zeng, & Hanson, 1996; Samejima, 1994), but none has yet become a test industry standard.

Temporal Stability

Are test scores consistent over time? Test scores must be reasonably consistent to have practical utility for making clinical and educational decisions and to be predictive of future performance. The stability coefficient, or test-retest score reliability coefficient, is an index of temporal stability that can be calculated by correlating test performance for a large number of examinees at two points in time. Two weeks is considered a preferred test-retest time interval (Nunnally & Bernstein, 1994; Salvia, Ysseldyke, & Bolt, 2010), because longer intervals increase the amount of error (due to maturation and learning) and tend to lower the estimated reliability. Because test-retest reliability and internal consistency forms of reliability are affected by different sources of error, it is possible for one to be high while the other is not (e.g., Nunnally & Bernstein, 1994).

Bracken (1987) recommended that a total test stability coefficient should be greater than or equal to .90 for high-stakes tests over relatively short test-retest intervals, whereas a stability coefficient of .80 is reasonable for low-stakes testing. Stability coefficients may be spuriously high, even with tests with low internal consistency, but tests with low stability coefficients tend to have low internal consistency unless they are tapping highly variable state-based constructs such as state anxiety (Nunnally & Bernstein, 1994). As a general rule of thumb, measures of internal consistency are preferred to stability coefficients as indexes of reliability.

Interrater Consistency and Consensus

Whenever tests require observers to render judgments, ratings, or scores for a specific behavior or performance, the consistency among observers constitutes an important source of measurement precision. Two separate methodological approaches have been utilized to study consistency and consensus among observers: interrater reliability (using correlational indexes to reference consistency among observers) and interrater agreement (addressing percent agreement among observers; e.g., Tinsley & Weiss, 1975). These distinctive approaches are necessary because it is possible to have high interrater reliability with low manifest agreement among raters if ratings are different but proportional. Similarly, it is possible to have low interrater reliability with high manifest agreement among raters if consistency indexes lack power because of restriction in range.

Interrater reliability refers to the proportional consistency of variance among raters and tends to be correlational. The simplest index involves correlation of total scores generated by separate raters. The *intraclass correlation* is another index of reliability commonly used to estimate the reliability of ratings. Its value ranges from 0 to 1.00, and it can be used to estimate the expected reliability of either the individual ratings provided by a single rater or the mean rating provided by a group of raters (Shrout & Fleiss, 1979). Another index of reliability, Kendall's *coefficient of concordance,* establishes how much reliability exists among ranked data. This procedure is appropriate when raters are asked to rank order the persons or behaviors along a specified dimension.

Interrater agreement refers to the interchangeability of judgments among raters, addressing the extent to which raters make the same ratings. Indexes of interrater agreement typically estimate percentage of agreement on categorical and rating decisions among observers, differing in the extent to which they are sensitive to degrees of agreement correct for chance agreement. Neuendorf (2002) reviewed rules of thumb proposed by a variety of researchers and concluded that "coefficients of .90 or greater would be acceptable to all, .80 or greater would be acceptable in most situations, and below that, there exists great disagreement" (p. 145). The criterion of .70 is often used for exploratory research. More liberal criteria are usually used for the indices known to be more conservative. An example is Cohen's kappa, a widely used statistic of interobserver agreement intended for situations in which raters classify the items being rated into discrete, nominal categories. Kappa ranges from −1.00 to +1.00; kappa values of .75 or higher are generally taken to indicate excellent agreement beyond chance; values between .60 and .74 are considered good agreement; those between .40 and .59 are considered fair; and those below .40 are considered poor (Fleiss, 1981).

Interrater reliability and agreement may vary logically depending on the degree of consistency expected from specific sets of raters. For example, it might be anticipated that people who rate a child's behavior in different contexts (e.g., school versus home) would produce lower correlations than two raters who rate the child within the same context (e.g., two parents within the home or two teachers at school). In a review of 13 preschool social-emotional instruments, the vast majority of reported coefficients of interrater congruence were below .80 (range .12–.89). Walker and Bracken (1996) investigated the congruence of biological parents who rated their children on four preschool behavior rating scales. Interparent congruence ranged from a low of .03 (Temperamental Assessment Battery for Children [TABC], Ease of Management through Distractibility) to a high of .79 (TABC, Approach/Withdrawal). In addition to concern about low congruence coefficients, the authors voiced concern that 44% of the parent pairs had a mean discrepancy across scales of 10 to 13 standard score points; differences ranged from 0 to 79 standard score points.

Interrater reliability studies are preferentially conducted under field conditions to enhance generalizability of testing by clinicians "performing under the time constraints and conditions of their work" (Wood, Nezworski, & Stejskal, 1996, p. 4). Cone (1988) described interscorer studies as fundamental to measurement because without scoring consistency and agreement, many other reliability and validity issues cannot be addressed.

Lombard, Snyder-Duch, and Bracken (2002) recommended that interrater reliability be reported with this information at minimum:

- Size of the reliability sample
- Method for selection of the sample
- Description of the relationship of the reliability sample to the full sample
- Number of and identity of the reliability raters

- Approximate training time required to reach adequate reliability
- Amount of coding completed by each rater
- How rating disagreements were resolved
- Indices selected to calculate reliability with a justification
- Interrater reliability for each variable selected
- Where and how consumers can obtain detailed information about the coding instrument, procedures, and instructions

Congruence Between Alternate Forms

When two parallel forms of a test are available, correlating scores on each form provides another way to assess reliability. In CTT, strict parallelism between forms requires equality of means, variances, and covariances (Gulliksen, 1950). A hierarchy of methods for pinpointing sources of measurement error with alternative forms has been proposed (Nunnally & Bernstein, 1994; Salvia et al., 2010): (1) assess alternate-form reliability with a 2-week interval between forms; (2) administer both forms on the same day; and, if necessary, (3) arrange for different raters to score the forms administered with a 2-week retest interval and on the same day. If the score correlation over the 2-week interval between the alternative forms is lower than coefficient alpha by .20 or more, considerable measurement error is present due to internal consistency, scoring subjectivity, or trait instability over time. If the score correlation is substantially higher for forms administered on the same day, the error may stem from trait variation over time. If the correlations remain low for forms administered on the same day, the two forms may differ in content with one form being more internally consistent than the other. If trait variation and content differences have been ruled out, comparison of subjective ratings from different sources may permit the major source of error to be attributed to the subjectivity of scoring.

In IRT, test forms may be compared by examining the forms at the item level. Forms with items of comparable item difficulties, response ogives, and standard errors by trait level will tend to have adequate levels of alternate form reliability (e.g., McGrew & Woodcock, 2001). For example, when item difficulties for one form are plotted against those for the second form, a clear linear trend is expected. When raw scores are plotted against trait levels for the two forms on the same graph, the ogive plots should be identical.

At the same time, scores from different tests tapping the same construct need not be parallel if both involve sets of items that are close to the examinee's ability level. As reported by Embretson (1999), "Comparing test scores across multiple forms is optimal when test difficulty levels vary across persons" (p. 12). The capacity of IRT to estimate trait level across differing tests does not require assumptions of parallel forms or test equating.

Reliability Generalization

In recognition that reliability is not inherent to a test itself and is influenced by test score hetereogeneity within a given sample, Vacha-Haase (1998) proposed *reliability generalization* as a meta-analytic methodology that investigates the reliability of scores across samples, studies, and administrative conditions. An extension of validity generalization (Hunter & Schmidt, 1990; Schmidt & Hunter, 1977), reliability generalization investigates the stability and variability of reliability coefficients across samples and studies and has now been reported for a number of measurements (see, e.g., Thompson, 2003). In order to demonstrate measurement precision for the populations for which a test is intended, a test should show comparable levels of reliability across various demographic subsets of the population (e.g., gender, race, ethnic groups) as well as salient clinical and exceptional populations. It is now considered best practice to report score reliabilities with CIs in recognition of the variability that may be found in test precision across samples (e.g., Fan & Thompson, 2001; Meyer, 2010).

TEST SCORE FAIRNESS

From the inception of psychological testing, concerns about fairness and potential bias have been apparent. As early as 1911, Alfred Binet (Binet & Simon, 1911/1916) was aware that a failure to represent diverse classes of SES would affect normative performance on intelligence tests. He intentionally deleted classes of items that related more to quality of education than to mental faculties. Early editions of the Stanford-Binet and the Wechsler intelligence scales were standardized on entirely white, native-born samples (Terman, 1916; Terman & Merrill, 1937; Wechsler, 1939, 1946, 1949). In addition to sample limitations, early tests also contained items that reflected positively on whites. Early editions of the Stanford-Binet included an Aesthetic Comparisons item in which examinees were shown a white, well-coiffed, blond woman and a disheveled woman with African features; the examinee was asked "Which one is prettier?" The original MMPI (Hathaway & McKinley, 1943) was normed on a convenience sample of White adult Minnesotans and contained items referring to culture-specific games ("drop-the-handkerchief"), literature (*Alice's Adventures in Wonderland*), and religious beliefs (the "second coming of Christ"). Most contemporary test developers now routinely avoid such problems with normative samples without minority representation as well as racially and ethnically insensitive items.

In spite of these advances, the fairness of educational and psychological tests represents one of the most contentious and psychometrically challenging aspects of test development. The examination of test psychometric properties across various groups, including majority and minority groups, may be considered a special form of test score validity. Numerous methodologies have been proposed to assess item and test properties for different groups of test takers, and the definitive text in this area is Jensen's (1980) thoughtful *Bias in Mental Testing*. Most of the controversy regarding test fairness relates to the legal, political, and social perceptions that group differences in test scores, or differences in selection rates, constitute evidence of bias in and of itself. For example, Jencks and Phillips (1998) stressed that the test score gap is the single most important obstacle to achieving racial balance and social equity.

In landmark litigation, Judge Robert Peckham in *Larry P. v. Riles* (343 F. Supp. 306, 1972; 495 F. Supp. 926, 1979) banned the use of individual IQ tests in placing black children into educable mentally retarded classes in California, concluding that the cultural bias of the IQ test was hardly disputed in this litigation. He asserted, "Defendants do not seem to dispute the evidence amassed by plaintiffs to demonstrate that the IQ tests in fact are culturally biased" (Peckham, 1972, p. 1313) and later concluded, "An unbiased test that measures ability or potential should yield the same pattern of scores when administered to different groups of people" (pp. 954–955).

The belief that any group test score difference constitutes bias has been termed the *egalitarian fallacy* by Jensen (1980):

> This concept of test bias is based on the gratuitous assumption that all human populations are essentially identical or equal in whatever trait or ability the test purports to measure. Therefore, any difference between populations in the distribution of test scores (such as a difference in means, or standard deviations, or any other parameters of the distribution) is taken as evidence that the test is biased. The search for a less biased test, then, is guided by the criterion of minimizing or eliminating the statistical differences between groups. The perfectly nonbiased test, according to this definition, would reveal reliable individual differences but not reliable (i.e., statistically significant) group differences. (p. 370)

However this controversy is viewed, the perception of test bias stemming from group mean score differences remains a deeply ingrained belief among many psychologists and educators. McArdle (1998) suggested that large group mean score differences are "a necessary but not sufficient condition for test bias" (p. 158). McAllister (1993) has observed that "[i]n the testing community, differences in correct answer rates, total scores, and so on do not mean bias. In the political realm, the exact opposite perception is found; differences mean bias" (p. 394).

The newest models of test fairness describe a systemic approach utilizing both internal and external sources of evidence of fairness that extend from test conception and design through test score interpretation and application (Camilli & Shepard, 1994; McArdle, 1998; Willingham, 1999). These models are important because they acknowledge the importance of the consequences of test use in a holistic assessment of fairness and a multifaceted methodological approach to accumulate evidence of test fairness. This section describes a systemic model of test fairness adapted from the work of several leading authorities.

Terms and Definitions

Three key terms appear in the literature associated with test score fairness: *bias, fairness,* and *equity*. These concepts overlap but are not identical; for example, a test that shows no evidence of test score bias may be used unfairly. To some extent these terms have historically been defined by families of relevant psychometric analyses—for example, bias is usually associated with differential item functioning, and fairness is associated with differential prediction to an external criterion. In this section, the terms are defined at a conceptual level.

Test score bias tends to be defined in a narrow manner, as a special case of test score invalidity. According to the most recent *Standards for Educational and Psychological Testing* (1999), bias in testing refers to "construct under-representation or construct-irrelevant components of test scores that differentially affect the performance of different groups of test takers" (p. 172). This definition implies that bias stems from nonrandom measurement error, provided that the typical magnitude of random error is comparable for all groups of interest. Accordingly, test score bias refers to the systematic and invalid introduction of measurement error for a particular group of interest. The statistical underpinnings of this definition have been underscored by Jensen (1980), who asserted, "The assessment of bias is a purely objective, empirical, statistical and quantitative matter entirely independent of subjective value judgments and ethical issues concerning fairness or unfairness of tests and the uses to which they are put" (p. 375). Some scholars consider the characterization of bias as objective and independent of the value judgments associated with fair use of tests to be fundamentally incorrect (e.g., Willingham, 1999).

Test score fairness refers to the ways in which test scores are utilized, most often for various forms of consequential decision making, such as selection or placement. Jensen (1980) suggested that the term refers "to the ways in which test scores (whether of biased or unbiased tests) are used in any selection situation" (p. 376), arguing that fairness is a subjective policy decision based on philosophic, legal, or practical considerations rather than a statistical decision. Willingham (1999) described a test fairness manifold that extends throughout the entire process of test development including the consequences of test usage. Embracing the idea that fairness is akin to demonstrating the generalizability of test validity across population subgroups, Willingham noted that "the manifold of fairness issues is complex because validity is complex" (p. 223). Fairness is a concept that transcends a narrow statistical and psychometric approach.

Finally, *equity* refers to a social value associated with the intended and unintended consequences and impact of test score usage. Because of the importance of equal opportunity, equal protection, and equal treatment in mental health, education, and the workplace, Willingham (1999) recommended that psychometrics actively consider equity issues in test development. As Tiedeman (1978) noted, "Test equity seems to be emerging as a criterion for test use on a par with the concepts of reliability and validity" (p. xxviii). However, the expectation that tests can correct long-standing problems of equity in society has never been grounded in psychometric science.

Internal Evidence of Fairness

The demonstration that a test has *equal internal integrity* across racial and ethnic groups has been described as a way to demonstrate test fairness (e.g., Mercer, 1984). The *internal* features of a test related to fairness generally include the test's theoretical underpinnings, item content and format, differential item and test functioning, reliability generalization, and measurement invariance. The two best-known procedures for evaluating test fairness include expert reviews of content bias and analysis of differential item functioning. This section discusses these and several additional sources of evidence of test fairness.

Item Bias and Sensitivity Review

In efforts to enhance fairness, the content and format of psychological and educational tests commonly undergo subjective bias and sensitivity reviews one or more times during test development. In this review, independent representatives from diverse groups closely examine tests, identifying items and procedures that may yield differential responses for one group relative to another. Content may be reviewed for cultural, disability, ethnic, racial, religious, sex, and SES bias. For example, a reviewer may be asked a series of questions, including "Does the content, format, or structure of the test item present greater problems for students from some backgrounds than for others?" A comprehensive item bias review is available from Hambleton and Rodgers (1995), and useful guidelines to reduce bias in language are available from the American Psychological Association (1994).

Ideally there are two objectives in bias and sensitivity reviews: (1) eliminate biased material and (2) ensure balanced and neutral representation of groups within the test. Among the potentially biased elements of tests that should be avoided is

- material that is controversial, or emotionally charged, or inflammatory for any specific group;
- language, artwork, or material that is demeaning or offensive to any specific group;
- content or situations with differential familiarity and relevance for specific groups;
- language and instructions that have different or unfamiliar meanings for specific groups;
- information and/or skills that may not be expected to be within the educational background of all examinees; and
- format or structure of the item that presents differential difficulty for specific groups.

Among the prosocial elements that ideally should be included in tests are

- presentation of universal experiences in test material;
- balanced distribution of people from diverse groups;
- presentation of people in activities that do not reinforce stereotypes;
- item presentation in a sex-, culture-, age-, and race-neutral manner; and
- inclusion of individuals with disabilities or handicapping conditions.

In general, the content of test materials should be relevant and accessible for the entire population of examinees for whom the test is intended. For example, the experiences of snow and freezing winters are outside the range of knowledge of many Southern students, thereby introducing a potential geographic regional bias. Use of utensils such as forks may be unfamiliar to Asian immigrants who may instead use chopsticks. Use of coinage from the United States ensures that the test cannot be validly used with examinees from countries with different currency.

Tests should also be free of controversial, emotionally charged, or value-laden content, such as violence or religion. The presence of such material may prove distracting, offensive, or unsettling to examinees from some groups, detracting from test performance. For example, Lichtenberger and Kaufman (2009) documented the removal of emotionally evocative and clinically rich items from the Wechsler intelligence scales—for example, *beer-wine* from the Similarities subtest and *knife* and *gamble* from the Vocabulary subtest—because these items elicited responses that sometimes tapped psychological constructs irrelevant to intelligence per se.

Stereotyping refers to the portrayal of a group using only a limited number of attributes, characteristics, or roles. As a rule, stereotyping should be avoided in test development. Specific groups should be portrayed accurately and fairly, without reference to stereotypes or traditional roles regarding sex, race, ethnicity, religion, physical ability, or geographic setting. Group members should be portrayed as exhibiting a full range of activities, behaviors, and roles.

Differential Item and Test Functioning

Are item and test statistical properties equivalent for individuals of comparable ability, but from different groups? *Differential test and item functioning* (DTIF, or DTF and DIF) refers to a family of statistical procedures aimed at determining whether examinees of the same ability but from different groups have different probabilities of success on a test or an item. The most widely used of DIF procedures is the Mantel-Haenszel technique (Holland & Thayer, 1988), which assesses similarities in item functioning across various demographic groups of comparable ability. Items showing significant DIF are usually considered for deletion from a test.

DIF has been extended by Shealy and Stout (1993) to a test score–based level of analysis known as differential test functioning, a multidimensional nonparametric IRT index of test bias. Whereas DIF is expressed at the item level, DTF represents a combination of two or more items to produce DTF, with scores on a valid subtest used to match examinees according to ability level. Tests may show evidence of DIF on some items without evidence of DTF, provided item bias statistics are offsetting and eliminate differential bias at the test score level.

Although psychometricians have embraced DIF as a preferred method for detecting potential item bias (McAllister, 1993), this methodology has been subjected to increasing criticism because of its dependence on internal test properties and its inherent circular reasoning. Hills (1999) noted that two decades of DIF research have failed to demonstrate that removing biased items affects test bias and narrows the gap in group mean scores. Furthermore, DIF rests on several assumptions including that items are unidimensional, that the latent trait is equivalently distributed across groups, that the groups being compared (usually racial, sex, or ethnic groups) are homogeneous, and that the overall test is unbiased. Camilli and Shepard (1994) observed: “By definition, internal DIF methods are incapable of detecting constant bias. Their aim, and capability, is only to detect relative discrepancies” (p. 17). Hunter and Schmidt (2000) have criticized DIF methodology, finding that most evidence of DIF may be explained by a failure to control for measurement error in ability estimates, violations of the DIF unidimensionality assumption, and/or reliance on spurious artifactual findings from statistical significance testing. Disparaging DIF methodology, they wrote, “[W]e know that the literature on item bias is unsound from a technical standpoint” (p. 157).

Measurement Invariance

Measurement invariance in psychological measurement is concerned with systematic group consistency in the information provided by a test about the latent variable or variables to be measured. Although there are multiple potential ways to demonstrate measurement invariance, the demonstration of *factorial invariance* across racial and ethnic groups via confirmatory factor analyses is one way to provide evidence that a test’s internal structure is invariant. Millsap (1997) observed, “When measurement bias is present, two individuals from different groups who are identical on the latent variable(s) of interest would be expected to score differently on the test” (p. 249).

A difference in the factor structure across groups provides some evidence for bias even though factorial invariance does not necessarily signify fairness (e.g., Meredith, 1993; Nunnally & Bernstein, 1994). Floyd and Widaman (1995) suggested that “[i]ncreasing recognition of cultural, developmental, and contextual influences on psychological constructs has raised interest in demonstrating measurement invariance before assuming that measures are equivalent across groups" (p. 296).

Reliability Generalization

Earlier we described the methodology of reliability generalization as a meta-analytic methodology that can investigate test score reliability across samples (Vacha-Haase, 1998). Reliability generalization also has the capacity to provide evidence of test score fairness by demonstrating relatively little change in score reliabilities across racial, ethnic, linguistic, and gender subsamples. Demonstration of adequate measurement precision across groups suggests that a test has adequate accuracy for the populations in which it may be used. Geisinger (1998) noted that

> subgroup-specific reliability analysis may be especially appropriate when the reliability of a test has been justified on the basis of internal consistency reliability procedures (e.g., coefficient *alpha*). Such analysis should be repeated in the group of special test takers because the meaning and difficulty of some components of the test may change over groups, especially over some cultural, linguistic, and disability groups. (p. 25)

Differences in group reliabilities may be evident, however, when test items are substantially more difficult for one group than another or when ceiling or floor effects are present for only one group.

The temporal stability of test scores should also be compared across groups, using similar test-retest intervals, in order to ensure that test results are equally stable irrespective of race and ethnicity. Jensen (1980) suggested:

> If a test is unbiased, test-retest correlation, of course with the same interval between testings for the major and minor groups, should yield the same correlation for both groups. Significantly different test-retest correlations (taking proper account of possibly unequal variances in the two groups) are indicative of a biased test. Failure to understand instructions, guessing, carelessness, marking answers haphazardly, and the like, all tend to lower the test-retest correlation. If two groups differ in test-retest correlation, it is clear that the test scores are not equally accurate or stable measures of both groups. (p. 430)

External Evidence of Fairness

Beyond the concept of internal integrity, Mercer (1984) recommended that studies of test fairness include evidence of *equal external relevance*. In brief, this determination requires the examination of relations between item/test scores and independent external criteria. External evidence of test score fairness has been accumulated in the study of comparative prediction of future performance (e.g., use of the SAT across racial groups to predict a student's ability to do college-level work). Fair prediction and fair selection are two objectives that are particularly important as evidence of test fairness, in part because they figure prominently in legislation and court rulings.

Fair Prediction

Prediction bias can arise when a test differentially predicts future behaviors or performance across groups. Cleary (1968) introduced a methodology that evaluates comparative predictive validity between two or more salient groups. The Cleary rule states that a test may be considered fair if it has the same approximate regression equation (i.e., comparable slope and intercept) explaining the relationship between the predictor test and an external criterion measure in the groups undergoing comparison. A slope difference between the two groups conveys differential validity and relates that one group's performance on the external criterion is predicted less well than the other's performance. An intercept difference suggests a difference in the level of estimated performance between the groups, even if the predictive validity is comparable. It is important to note that this methodology assumes adequate levels of reliability for both the predictor and criterion variables. This procedure has several limitations that have been summarized by Camilli and Shepard (1994). The demonstration of equivalent predictive validity across demographic groups constitutes an important source of fairness that is related to validity generalization. Millsap (1997) demonstrated, however, that measurement invariance may be logically and statistically incompatible with invariant prediction, challenging the value of conventional approaches to prediction bias. Hunter and Schmidt (2000) went further in their appraisal of the fair prediction literature:

> For the past 30 years, civil rights lawyers, journalists, and others . . . have argued that when test scores are equal, minorities have higher average levels of educational and work performance, meaning that test scores underestimate the real world performance of minorities. Thousands of test bias studies have been conducted, and these studies have disconfirmed that hypothesis. The National Academy of Science . . . concluded that professionally developed tests are not predictively biased. (p. 151)

Based on these observations pertaining to predictive bias, Hunter and Schmidt concluded that "the issue of test bias is scientifically dead" (p. 151).

Fair Selection

The consequences of test score use for selection and decision making in clinical, educational, and occupational domains constitute a source of potential bias. The issue of fair selection addresses the question: Do the use of test scores for selection decisions unfairly favor one group over another? Specifically, test scores that produce adverse, disparate, or disproportionate impact for various racial or ethnic groups may be said to show evidence of selection bias, even when that impact is construct relevant. Since enactment of the Civil Rights Act of 1964, demonstration of adverse impact has been treated in legal settings as prima facie evidence of test bias. Adverse impact occurs when there is a substantially different rate of selection based on test scores and other factors works to the disadvantage of members of a race, sex, or ethnic group.

Federal mandates and court rulings often have indicated that adverse, disparate, or disproportionate impact in selection decisions based on test scores constitutes evidence of unlawful discrimination, and differential test selection rates among majority and minority groups have been considered a bottom line in federal mandates and court rulings. In its Uniform Guidelines on Employment Selection Procedures (1978), the Equal Employment Opportunity Commission operationalized adverse impact according to the four-fifths rule, which states: "A selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact" (p. 126). Adverse impact has been applied to educational tests (e.g., the *Texas Assessment of Academic Skills*) as well as tests used in personnel selection. The U.S. Supreme Court held in 1988 that differential selection ratios can constitute sufficient evidence of adverse impact. The 1991 Civil Rights Act, Section 9, specifically and explicitly prohibits any discriminatory use of test scores for minority groups.

Since selection decisions involve the use of test cutoff scores, an analysis of costs and benefits according to decision theory provides a methodology for fully understanding the consequences of test score usage. Cutoff scores may be varied to provide optimal fairness across groups, or alternative cutoff scores may be utilized in certain circumstances. McArdle (1998) observed, "As the cutoff scores become increasingly stringent, the number of false negative mistakes (or costs) also increase, but the number of false positive mistakes (also a cost) decrease" (p. 174).

LIMITS OF PSYCHOMETRICS

Psychological assessment is ultimately about the examinee. A test is merely a tool to understand the examinee, and psychometrics are merely rules to build and evaluate the tools. The tools themselves must be sufficiently sound (i.e., valid, reliable) and fair so that they introduce acceptable levels of error into the process of decision making. Some of the guidelines that have been described in this chapter for psychometrics of test construction and application help us not only to build better tools but to use these tools as skilled craftspersons.

As an evolving field of study, psychometrics still has some glaring shortcomings. A long-standing limitation of psychometrics is its systematic overreliance on internal sources of evidence for test validity and fairness. In brief, it is more expensive and more difficult to collect external criterion-based information, especially with special populations; it is simpler and easier to base all analyses on the performance of a normative standardization sample. This dependency on internal methods has been recognized and acknowledged by leading psychometricians. In discussing psychometric methods for detecting test bias, for example, Camilli and Shepard (1994) cautioned about circular reasoning: "Because DIF indices rely only on internal criteria, they are inherently circular" (p. 17). Similarly, psychometricians have been hesitant to consider attempts to extend the domain of validity into consequential aspects of test usage (e.g., Borsboom 2006a; Lees-Haley, 1996). We have witnessed entire testing approaches based on internal factor-analytic approaches and evaluation of content validity (e.g., McGrew & Flanagan, 1998), with comparatively little attention paid to the external validation of the factors against independent, ecologically valid criteria. This shortcoming constitutes a serious limitation of psychometrics, which we have attempted to address by encouraging use of both internal and external sources of psychometric evidence.

Another long-standing limitation is the tendency of test developers to wait until the test is undergoing standardization to establish its validity through clinical studies. A typical sequence of test development involves pilot studies, a content tryout, and finally a national standardization and supplementary studies (e.g., Robertson, 1992). In the stages of test development described by Loevinger (1957), the external criterion-based validation stage comes last in the process—after the test has effectively been built. A limitation in test development and psychometric practice is that many tests validate their effectiveness for a stated purpose only at the end of the process rather than at the beginning, as MMPI developers did over a half century ago by selecting items that discriminated between specific diagnostic groups (Hathaway & McKinley, 1943). The utility of a test for its intended application should be at least partially validated at the pilot study stage, prior to norming. Even better is an evidence-based validity argument, such as proposed by Mislevy and Haertel (2006), to explicitly link test construction with its intended application.

Finally, psychometrics has failed to directly address many of the applied questions of practitioners. Test results often do not readily lend themselves to functional decision making. For example, psychometricians have been slow to develop consensually accepted ways of measuring growth and maturation, reliable change (as a result of enrichment, intervention, or treatment), and atypical response patterns suggestive of lack of effort or dissimilation. There is a strong need for more innovations like the Bayesian nomogram, which readily lends itself to straightforward clinical application (e.g., Bianchi, Alexander, & Cash, 2009; Jenkins et al., 2011). In general, the failure of treatment validity and assessment–treatment linkage undermines the central purpose of testing.

Looking to the future, the emergence of evidence-based assessment (EBA) guidelines now appears inevitable, paralleling the numerous evidence-based treatment, instruction, and intervention effectiveness studies that have led to professional practice guidelines. While EBA has taken many different forms in the literature, Hunsley and Mash (2005) have anticipated that it will include standard psychometric indices of reliability and validity and also encompass treatment utility, diagnostic utility, and a range of additional factors, such as the economic and psychological costs associated with assessment error. Any forthcoming rules for EBA are likely to increase the accountability of test users (who will increasingly be required to use empirically supported measurements) and test developers (who will need to carefully determine the new mix of psychometric studies necessary to successfully meet professional needs). The powerful impact of a single psychometrically oriented review—Lilienfeld, Wood, and Garb's (2000) critical assessment of projective tests—may provide a hint of the magnitude of changes that may come with EBA.

REFERENCES

Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA school-age forms and profiles.* Burlington: University of Vermont, Research Center for Children, Youth, & Families.

American Psychological Association. (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, DC: Author.

American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist, 57,* 1060–1073.

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

Andrich, D. (1988). *Rasch models for measurement.* Thousand Oaks, CA: Sage.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores.* Princeton, NJ: Educational Testing Service.

Banaji, M. R., & Crowder, R. C. (1989). The bankruptcy of everyday memory. *American Psychologist, 44,* 1185–1193.

Barrios, B. A. (1988). On the changing nature of behavioral assessment. In A. S. Bellack & M. Hersen (Eds.), *Behavioral assessment: A practical handbook* (3rd ed., pp. 3–41). New York, NY: Pergamon Press.

Bayley, N. (1993). *Bayley Scales of Infant Development, Second edition manual.* San Antonio, TX: Psychological Corporation.

Ben-Porath, Y. S., & Tellegen, A. (2008). *MMPI-2-RF: Manual for administration, scoring, and interpretation.* Minneapolis: University of Minnesota Press.

Beutler, L. E. (1998). Identifying empirically supported treatments: What if we didn't? *Journal of Consulting & Clinical Psychology, 66,* 113–120.

Beutler, L. E., & Clarkin, J. F. (1990). *Systematic treatment selection: Toward targeted therapeutic interventions.* Philadelphia, PA: Brunner/Mazel.

Beutler, L. E., & Harwood, T. M. (2000). *Prescriptive psychotherapy: A practical guide to systematic treatment selection.* New York, NY: Oxford University Press.

Bianchi, M. T., Alexander, B. M., & Cash, S. S. (2009). Incorporating uncertainty into medical decision making: An approach to unexpected test results. *Medical Decision Making, 29,* 116–124.

Binet, A., & Simon, T. (1911/1916). New investigation upon the measure of the intellectual level among school children. *L'Année Psychologique, 17,* 145–201. In E. S. Kite (Trans.), *The development of intelligence in children* (pp. 274–329). Baltimore, MD: Williams & Wilkins. (Original work published 1911)

Borsboom, D. (2006a). The attack of the psychometricians. *Psychometrika, 71,* 425–440.

Borsboom, D. (2006b). Can we bring out a velvet revolution in psychological measurement? A rejoinder to commentaries. *Psychometrika, 71,* 463–467.

Borsboom, D., Cramer, A. O. J., Kievit, R. A., Scholten, A. Z., & Franić, S. (2009). The end of construct validity. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 135–170). Charlotte, NC: Information Age.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111,* 1061–1071.

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. (1988). Ten psychometric reasons why similar tests produce dissimilar results. *Journal of School Psychology, 26,* 155–166.

Bracken, B. A., & McCallum, R. S. (1998). *Universal Nonverbal Intelligence Test examiner's manual.* Itasca, IL: Riverside.

Brennan, R. L. (2010a). Evidence-centered assessment design and the Advanced Placement Program: A psychometrician's perspective. *Applied Measurement in Education, 23,* 392–400.

Brennan, R. L. (2010b). *Generalizability theory.* New York, NY: Springer-Verlag.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3,* 296–322.

Bruininks, R. H., Woodcock, R. W., Weatherman, R. F., & Hill, B. K. (1996). *Scales of Independent Behavior—Revised comprehensive manual.* Itasca, IL: Riverside.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items* (Vol. 4). Thousand Oaks, CA: Sage.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago, IL: Rand McNally.

Campbell, S. K., Siegel, E., Parr, C. A., & Ramey, C. T. (1986). Evidence for the need to renorm the Bayley Scales of Infant Development based on the performance of a population-based sample of 12-month-old infants. *Topics in Early Childhood Special Education, 6,* 83–96.

Carroll, J. B. (1983). Studying individual differences in cognitive abilities: Through and beyond factor analysis. In R. F. Dillon & R. R. Schmeck (Eds.), *Individual differences in cognition* (pp. 1–33). New York, NY: Academic Press.

Cattell, R. B. (1986). The psychometric properties of tests: Consistency, validity, and efficiency. In R. B. Cattell & R. C. Johnson (Eds.), *Functional psychological testing: Principles and instruments* (pp. 54–78). New York, NY: Brunner/Mazel.

Chudowsky, N., & Behuniak, P. (1998). Using focus groups to examine the consequential aspect of validity. *Educational Measurement: Issues & Practice, 17,* 28–38.

Cizek, G. J., Bowen, D., & Church, K. (2010). Sources of validity evidence for educational and psychological tests: A follow-up study. *Educational and Psychological Measurement, 70,* 732–743.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7,* 309–319.

Cleary, T. A. (1968). Test bias: Prediction of grades for Negro and White students in integrated colleges. *Journal of Educational Measurement, 5,* 115–124.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159.

Cone, J. D. (1978). The behavioral assessment grid (BAG): A conceptual framework and a taxonomy. *Behavior Therapy, 9,* 882–888.

Cone, J. D. (1988). Psychometric considerations and the multiple models of behavioral assessment. In A. S. Bellack & M. Hersen (Eds.), *Behavioral assessment: A practical handbook* (3rd ed., pp. 42–66). New York, NY: Pergamon Press.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago, IL: Rand McNally.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78,* 98–104.

Costa, P. T. Jr., & McCrae, R. R. (2010). *NEO Inventories Professional Manual for NEO-PI-3, NEO FFI-3 & NEO PIR.* Lutz, FL: Psychological Assessment Resources.

Crocker, L. (2003). Teaching for the test: Validity, fairness, and moral action. *Educational Measurement: Issues and Practice, 22,* 5–11.

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

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12,* 671–684.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions.* Urbana, IL: University of Illinois Press.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability scores and profiles.* New York, NY: Wiley.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Daniel, M. H. (1999). Behind the scenes: Using new measurement methods on the DAS and KAIT. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 37–63). Mahwah, NJ: Erlbaum.

Daniel, M. H. (2007). *Test norming.* Retrieved from www.pearsonassessments.com/pai/ca/RelatedInfo/TestNorming.htm

Dimitrov, D. M. (2002). Reliability: Arguments for multiple perspectives and potential problems with generalization across studies. *Educational and Psychological Measurement, 62,* 783–801.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* New York, NY: Chapman & Hall/CRC.

Elliott, C. D. (1990). *Differential Ability Scales: Introductory and technical handbook.* San Antonio, TX: Psychological Corporation.

Embretson, S. E. (1995). The new rules of measurement. *Psychological Assessment, 8,* 341–349.

Embretson, S. E. (1999). Issues in the measurement of cognitive abilities. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 1–15). Mahwah, NJ: Erlbaum.

Embretson, S. E. (2004). The second century of ability testing: Some predictions and speculations. *Measurement, 2,* 1–32.

Embretson, S. E., & Hershberger, S. L. (Eds.). (1999). *The new rules of measurement: What every psychologist and educator should know.* Mahwah, NJ: Erlbaum.

Ercikan, K., Arim, R., Law, D., Domene, J., Gagnon, F., & Lacroix, S. (2011). Application of Think Aloud Protocols for examining and confirming sources of differential item functioning identified by expert reviews. *Educational Measurement: Issues and Practice, 29*(2), 24–35.

Erdodi, L. A., Richard, D. C. S., & Hopwood, C. (2009). The importance of relying on the manual: Scoring error variance in the WISC-IV Vocabulary subtest. *Journal of Psychoeducational Assessment, 27,* 374–385.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87,* 215–251.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58,* 357–381.

Fan, X., & Thompson, B. (2001). Confidence intervals about score reliability coefficients, please: An EPM guidelines editorial. *Editorial and Psychological Measurement, 61,* 517–531.

Fiske, D. W., & Campbell, D. T. (1992). Citations do not solve problems. *Psychological Bulletin, 112,* 393–395.

Fleiss, J. L. (1981). Balanced incomplete block designs for inter-rater reliability studies. *Applied Psychological Measurement, 5,* 105–112.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7,* 286–299.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95,* 29–51.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101,* 171–191.

Flynn, J. R. (1994). IQ gains over time. In R. J. Sternberg (Ed.), *The encyclopedia of human intelligence* (pp. 617–623). New York, NY: Macmillan.

Flynn, J. R. (1999). Searching for justice: The discovery of IQ gains over time. *American Psychologist, 54,* 5–20.

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin, 137,* 316–344.

Foy, P., & Joncas, M. (2004). TIMSS 2003 sampling design. In M. O. Martin, I. V. S. Mullis, & S. J. Chrostowski (Eds.), *Trends in*

International Mathematics and Science Study (TIMSS) 2003 technical report (pp. 109–122). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Frisby, C. L., & Kim, S. (2008). Using profile analysis via multidimensional scaling (PAMS) to identify core profiles from the WMS-III. *Psychological Assessment, 20,* 1–9.

Galton, F. (1879). Psychometric experiments. *Brain: A Journal of Neurology, 2,* 149–162.

Geisinger, K. F. (1998). Psychometric issues in test interpretation. In J. Sandoval, C. L. Frisby, K. F. Geisinger, J. D. Scheuneman, & J. R. Grenier (Eds.) *Test interpretation and diversity: Achieving equity in assessment* (pp. 17–30). Washington, DC: American Psychological Association.

Gignac, G. E., Bates, T. C., & Jang, K. (2007). Implications relevant to CFA model misfit, reliability, and the five-factor model as measured by the NEO FFI. *Personality and Individual Differences, 43,* 1051–1062.

Gorsuch, R. L. (1983a). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Gorsuch, R. L. (1983b). The theory of continuous norming. In R. L. Gorsuch (Chair), *Continuous norming: An alternative to tabled norms?* Symposium conducted at the 91st Annual Convention of the American Psychological Association, Anaheim, CA.

Gorsuch, R. L. (2010). Continuous parameter estimation model: Expanding our statistical paradigm. In G. H. Roid (Chair), *Continuous parameter estimation and continuous norming methods.* Symposium conducted at the 118th annual convention of the American Psychological Association, San Diego, CA.

Gorsuch, R. L., & Zachary, R. A. (1985). Continuous norming: Implication for the WAIS-R. *Journal of Clinical Psychology, 41,* 86–94.

Guilford, J. P. (1950). *Fundamental statistics in psychology and education* (2nd ed.). New York, NY: McGraw-Hill.

Guion, R. M. (1977). Content validity—The source of my discontent. *Applied Psychological Measurement, 1,* 1–10.

Gulliksen, H. (1950). *Theory of mental tests.* New York, NY: McGraw-Hill.

Hambleton, R., & Rodgers, J. H. (1995). *Item bias review.* (ERIC Clearinghouse on Assessment and Evaluation, EDO-TM-95–9). Washington, DC: Catholic University of America Department of Education.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Hathaway, S. R. & McKinley, J. C. (1943). *Manual for the Minnesota Multiphasic Personality Inventory.* New York, NY: Psychological Corporation.

Hayes, S. C., Nelson, R. O., & Jarrett, R. B. (1987). The treatment utility of assessment: A functional approach to evaluating assessment quality. *American Psychologist, 42,* 963–974.

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7,* 238–247.

Heinrichs, R. W. (1990). Current and emergent applications of neuropsychological assessment problems of validity and utility. *Professional Psychology: Research and Practice, 21,* 171–176.

Helmes, E., & Reddon, J. R. (1993). A perspective on developments in assessing psychopathology: A critical review of the MMPI and MMPI-2. *Psychological Bulletin, 113,* 453–471.

Herrnstein, R. J., & Murray, C. (1994). *The bell curve: Intelligence and class in American life.* New York, NY: Free Press.

Hertzog, M. A. (2008). Considerations in determining sample size for pilot studies. *Research in Nursing & Health, 31,* 180–191.

Hills, J. (1999, May 14). Re: Construct validity. *Educational Statistics Discussion List (EDSTAT-L).* Available e-mail: edstat-l@jse.stat.ncsu.edu

Hoelzle, J. B., & Meyer, G. J. (2008). The factor structure of the MMPI–2 Restructured Clinical (RC) scales. *Journal of Personality Assessment, 90,* 443–455.

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement, 60,* 523–531.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

Horn, J. L., Wanberg, K. W., & Appel, M. (1973). On the internal structure of the MMPI. *Multivariate Behavioral Research, 8,* 131–171.

Hopkins, C. D., & Antes, R. L. (1978). *Classroom measurement and evaluation.* Itasca, IL: F. E. Peacock.

Hopwood, C., & Richard, D. C. S. (2005). WAIS-III scoring accuracy is a function of scale IC and complexity of examiner tasks. *Assessment, 12,* 445–454.

Hunsley, J., & Mash, E. J. (2005). Introduction to the special section on developing guidelines for the evidence-based assessment (EBA) of adult disorders. *Psychological Assessment, 17,* 251–255.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

Hunter, J. E., & Schmidt, F. L. (2000). Racial and gender bias in ability and achievement tests: Resolving the apparent paradox. *Psychology, Public Policy & Law, 6,* 151–158.

Hunter, J. E., Schmidt, F. L., & Jackson, C. B. (1982). *Advanced meta-analysis: Quantitative methods of cumulating research findings across studies.* San Francisco, CA: Sage.

Ittenbach, R. F., Esters, I. G., & Wainer, H. (1997). The history of test development. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 17–31). New York, NY: Guilford Press.

Jackson, D. N. (1971). A sequential system for personality scale development. In C. D. Spielberger (Ed.), *Current topics in clinical and community psychology* (Vol. 2, pp. 61–92). New York, NY: Academic Press.

Jencks, C., & Phillips, M. (Eds.). (1998). *The Black-White test score gap.* Washington, DC: Brookings Institute.

Jenkins, M. A., Youngstrom, E. A., Washburn, J. J., & Youngstrom, J. K. (2011). Evidence-based strategies improve assessment of pediatric bipolar disorder by community practitioners. *Professional Psychology: Research and Practice, 42,* 121–129.

Jensen, A. R. (1980). *Bias in mental testing.* New York, NY: Free Press.

Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika, 36,* 149–176.

Julious, S. A. (2005). Sample size of 12 per group rule of thumb for a pilot study. *Pharmaceutical Statistics, 4,* 287–291.

Kalton, G. (1983). *Introduction to survey sampling.* Beverly Hills, CA: Sage.

Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods, 11,* 363–385.

Kelley, T. L. (1927). *Interpretation of educational measurements.* New York, NY: Macmillan.

Knowles, E. S., & Condon, C. A. (2000). Does the rose still smell as sweet? Item variability across test forms and revisions. *Psychological Assessment, 12,* 245–252.

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement, 33,* 129–140.

Kuhn, T. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago, IL: University of Chicago Press.

Lazarus, A. A. (1973). Multimodal behavior therapy: Treating the BASIC ID. *Journal of Nervous and Mental Disease, 156,* 404–411.

Lees-Haley, P. R. (1996). Alice in validityland, or the dangerous consequences of consequential validity. *American Psychologist, 51,* 981–983.

Lehtonen, R., & Pahkinen, E. (2004). *Practical methods for design and analysis of complex surveys* (2nd ed.). Hoboken, NJ: Wiley.

Levy, P. S., & Lemeshow, S. (1999). *Sampling of populations: Methods and applications.* New York, NY: Wiley.

Li, H., Rosenthal, R., & Rubin, D. B. (1996). Reliability of measurement in psychology: From Spearman-Brown to maximal reliability. *Psychological Methods, 1,* 98–107.

Lichtenberger, E. O., & Kaufman, A. S. (2009). *Essentials of WAIS-IV assessment.* Hoboken, NJ: Wiley.

Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest, 1,* 27–66.

Linacre, J. M., & Wright, B. D. (1999). *A user's guide to Winsteps/Ministep: Rasch-model computer programs.* Chicago, IL: MESA Press.

Linn, R. L. (1998). Partitioning responsibility for the evaluation of the consequences of assessment programs. *Educational Measurement: Issues & Practice, 17,* 28–30.

Lissitz, R. W. (Ed.) (2009). *The concept of validity: Revisions, new directions, and applications.* Charlotte, NC: Information Age.

Loevinger, J. (1957). Objective tests as instruments of psychological theory [Monograph]. *Psychological Reports, 3,* 635–694.

Loevinger, J. (1972). Some limitations of objective personality tests. In J. N. Butcher (Ed.), *Objective personality assessment* (pp. 45–58). New York, NY: Academic Press.

Lohman, D. F., & Lakin, J. (2009). Consistencies in sex differences on the Cognitive Abilities Test across countries, grades, test forms, and cohorts. *British Journal of Educational Psychology, 79,* 389–407.

Loken, E., & Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology, 63,* 509–525.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research, 28,* 587–604.

Lord, F. N., & Novick, M. (1968). *Statistical theories of mental tests.* New York, NY: Addison-Wesley.

Maruish, M. E. (Ed.) (1994). *The use of psychological testing for treatment planning and outcome assessment.* Hillsdale, NJ: Erlbaum.

McAllister, P. H. (1993). Testing, DIF, and public policy. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 389–396). Hillsdale, NJ: Erlbaum.

McArdle, J. J. (1998). Contemporary statistical models for examining test-bias. In J. J. McArdle & R. W. Woodcock (Eds.), *Human cognitive abilities in theory and practice* (pp. 157–195). Mahwah, NJ: Erlbaum.

McCrae, R. R., Zonderman, A. B., Costa, P. T., Bond, M. H., & Paunonen, S. V. (1996). Evaluating replicability of factors in the revised NEO personality inventory: Confirmatory factor analysis versus Procrustes rotation. *Journal of Personality and Social Psychology, 70,* 552–566.

McGrew, K. S., Dailey, D. E. H., & Schrank, F. A. (2007). *Woodcock-Johnson III/Woodcock-Johnson III Normative Update score differences: What the user can expect and why (Woodcock-Johnson III Assessment Service Bulletin No. 9).* Rolling Meadows, IL: Riverside.

McGrew, K. S., & Flanagan, D. P. (1998). *The intelligence test desk reference (ITDR): Gf-Gc cross-battery assessment.* Boston, MA: Allyn & Bacon.

McGrew, K. S., & Woodcock, R. W. (2001). *Woodcock-Johnson III technical manual.* Itasca, IL: Riverside.

Mercer, J. R. (1984). What is a racially and culturally nondiscriminatory test? A sociological and pluralistic perspective. In C. R. Reynolds & R. T. Brown (Eds.), *Perspectives on bias in mental testing* (pp. 293–356). New York, NY: Plenum Press.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58,* 525–543.

Messick, S. (1989a). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18,* 5–11.

Messick, S. (1989b). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Washington, DC: American Council on Education and National Council on Measurement in Education.

Messick, S. (1995a). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues & Practice, 14,* 5–8.

Messick, S. (1995b). Validity of psychological assessment: Validation of inferences from persons' reponses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741–749.

Meyer, P. (2010). *Reliability.* New York, NY: Oxford University Press.

Millon, T., Davis, R., & Millon, C. (1997). *MCMI-III: Millon Clinical Multiaxial Inventory-III manual* (3rd ed.). Minneapolis, MN: National Computer Systems.

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods, 2,* 248–260.

Mislevy, R. J., & Haertel, G. (2006). Implications for evidence-centered design for educational assessment. *Educational Measurement: Issues and Practice, 25,* 6–20.

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A nonparametric approach to statistical inference.* Newbury Park, CA: Sage.

Neisser, U. (1978). Memory: What are the important questions? In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory* (pp. 3–24). London, UK: Academic Press.

Neuendorf, K. A. (2002). *The content analysis guidebook.* Thousand Oaks, CA: Sage.

Newborg, J., Stock, J. R., Wnek, L., Guidubaldi, J., & Svinicki, J. (1984). *Battelle Developmental Inventory.* Itasca, IL: Riverside.

No Child Left Behind Act of 2001. Pub. L. No. 107–110, 20 U.S.C. § 6301 *et seq.* (2002).

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Oakes, J. M., & Rossi, P. H. (2003). The measurement of SES in health research: current practice and steps toward a new approach. *Social Science & Medicine, 56,* 769–784.

Pearson, K. (1924). *The life, letters and labours of Francis Galton* (Vol. 2). *Researches of middle life.* London, UK: Cambridge University Press.

Peckham, R. F. (1972). Opinion, *Larry P.* v. *Riles*. *Federal Supplement 343,* 1306–1315.

Peckham, R. F. (1979). Opinion, *Larry P.* v. *Riles*. *Federal Supplement 495,* 926–992.

Pomplun, M. (1997). State assessment and instructional change: A path model analysis. *Applied Measurement in Education, 10,* 217–234.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danish Institute for Educational Research.

Reckase, M. D. (1998). Consequential validity from the test developer's perspective. *Educational Measurement: Issues & Practice, 17,* 13–16.

Reschly, D. J. (1997). Utility of individual ability measures and public policy choices for the 21st century. *School Psychology Review, 26,* 234–241.

Reschly, D. J., Myers, T. G., & Hartel, C. R. (2002). *Mental retardation: Determining eligibility for Social Security Benefits.* Washington, DC: National Academies Press.

Riese, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment, 12,* 287–297.

Robertson, G. J. (1992). Psychological tests: Development, publication, and distribution. In M. Zeidner & R. Most (Eds.), *Psychological testing: An inside view* (pp. 159–214). Palo Alto, CA: Consulting Psychologists Press.

Roid, G. (2003). *Stanford-Binet Intelligence Scales* (5th ed.). *Interpretive manual: Expanded guide to the interpretation of SB5 test results*. Itasca, IL: Riverside.

Roid, G. H. (2010). Update and new evidence for continuous norming. In G. H. Roid (Chair), *Continuous parameter estimation and continuous norming methods.* Symposium conducted at the 118th annual convention of the American Psychological Association, San Diego, CA.

Rust, K. F., & Johnson, E. G. (1992). Sampling and weighting in the National Assessment. *Journal of Educational Statistics, 17,* 111–129.

Salvia, J., Ysseldyke, J., & Bolt, S. (2010). *Assessment in special and inclusive education* (11th ed.). Belmont, CA: Wadsworth.

Samejima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement, 18,* 229–244.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62,* 529–540.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58,* 159–194.

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

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 19–37). Charlotte, NC: Information Age.

Slaney, K. L., & Maraun, M. D. (2008). A proposed framework for conducting data-based test analysis. *Psychological Methods, 13,* 376–390.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3,* 171–195.

Standards for educational and psychological testing. (1999). Washington, DC: American Educational Research Association.

Stigler, S. M. (1999). *Statistics on the table: The history of statistical concepts and methods.* Cambridge, MA: Harvard University Press.

Stigler, S. M. (2010). Darwin, Galton and the statistical enlightenment. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 173,* 469–482.

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80,* 99–103.

Suen, H. K. (1990). *Principles of test theories.* Hillsdale, NJ: Erlbaum.

Superfine, B. M. (2004). At the intersection of law and psychometrics: Explaining the validity clause of *No Child Left Behind*. *Journal of Law and Education, 33,* 475–514.

Swets, J. A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist, 47,* 522–532.

Teasdale, T. W., & Owen, D. R. (2005). A long-term rise and recent decline in intelligence test performance: The Flynn effect in reverse. *Personality and Individual Differences, 39,* 837–843.

Tellegen, A., & Ben-Porath, Y. S. (2008). *MMPI-2-RF (Minnesota Multiphasic Personality Inventory-2): Technical manual.* Minneapolis: University of Minnesota Press.

Tellegen, A., Ben-Porath, Y. S., McNulty, J. L., Arbisi, P. A., Graham, J. R., & Kaemmer, B. (2003). *The MMPI-2 Restructured Clinical (RC) scales: Development, validation, and interpretation.* Minneapolis: University of Minnesota Press.

Terman, L. M. (1916). *The measurement of intelligence: An explanation of and a complete guide for the use of the Stanford revision and extension of the Binet Simon Intelligence Scale.* Boston, MA: Houghton Mifflin.

Terman, L. M. & Merrill, M. A. (1937). *Directions for administering: Forms L and M, Revision of the Stanford-Binet Tests of Intelligence.* Boston, MA: Houghton Mifflin.

Thompson, B. (Ed.) (2003). *Score reliability: Contemporary thinking on reliability issues.* Thousand Oaks, CA: Sage.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* Washington, DC: American Psychological Association.

Tiedeman, D. V. (1978). In O. K. Buros (Ed.), *The eight mental measurements yearbook.* Highland Park: NJ: Gryphon Press.

Tinsley, H. E. A., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology, 22,* 358–376.

Tulsky, D. S., Chiaravallotti, N. D., Palmer, B. W., & Chelune, G. J. (2003). The Wechsler Memory Scale (3rd ed.): A new perspective. In D. S. Tulsky et al. (Eds.), *Clinical interpretation of the WAIS-III and WMS-III* (pp. 93–139). New York, NY: Academic Press.

Tulsky, D. S., & Ledbetter, M. F. (2000). Updating to the WAIS-III and WMS-III: Considerations for research and clinical practice. *Psychological Assessment, 12,* 253–262.

Tulsky, D. [S.], Zhu, J., & Ledbetter, M. F. (1997). *WAIS-III/WMS-III technical manual.* San Antonio, TX: Psychological Corporation.

Turkheimer, E., Haley, A., Waldron, M., D'Onofrio, B., & Gottesman, I. I. (2003). Socioeconomic status modifies heritability of IQ in young children. *Psychological Science, 14,* 623–628.

Uniform guidelines on employee selection procedures. (1978). *Federal Register, 43,* 38296–38309.

Urbina, S. (2004). *Essentials of psychological testing.* Hoboken, NJ: Wiley.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational & Psychological Measurement, 58,* 6–20.

Vacha-Haase, T., Kogan, L. R., & Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement, 60,* 509–522.

Vassend, O., & Skrondal, A. (2011). The NEO personality inventory revised (NEO-PI-R): Exploring the measurement structure and variants of the five-factor model. *Personality and Individual Differences, 50,* 1300–1304.

Vaughn, S., & Fuchs, L. S. (2003). Redefining learning disabilities as inadequate response to instruction: The promise and potential problems. *Learning Disabilities Research & Practice, 18*(3), 137–146.

Vogt, D. S., King, D. W., & King, L. A. (2004). Focus groups in psychological assessment: Enhancing content validity by consulting members of the target population. *Psychological Assessment, 16,* 231–243.

Wainer, H. (2010). 14 conversations about three things. *Journal of Educational and Behavioral Statistics, 35,* 5–25.

Walker, K. C., & Bracken, B. A. (1996). Inter-parent agreement on four preschool behavior rating scales: Effects of parent and child gender. *Psychology in the Schools, 33,* 273–281.

Wechsler, D. (1939). *The measurement of adult intelligence.* Baltimore, MD: Williams & Wilkins.

Wechsler, D. (1946). *The Wechsler-Bellevue intelligence scale: Form II. Manual for administering and scoring the test.* New York, NY: Psychological Corporation.

Wechsler, D. (1949). *Wechsler Intelligence Scale for Children manual.* New York, NY: Psychological Corporation.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children* (3rd ed.). San Antonio, TX: Psychological Corporation.

Wechsler, D. (1997). *Wechsler Memory Scale—Third Edition (WMS-III) administration and scoring manual.* San Antonio, TX: Psychological Corporation.

Wechsler, D. (2003). *WISC-IV technical and interpretive manual.* San Antonio, TX: Psychological Corporation.

Wechsler, D. (2008). *WAIS-IV technical and interpretive manual.* San Antonio, TX: Pearson.

Weed, N. C. (2006). Syndrome complexity, paradigm shifts, and the future of validation research: Comments on Nichols and Rogers, Sewell, Harrison, and Jordan. *Journal of Personality Assessment, 87,* 217–222.

Wilkins, C., & Rolfhus, E. (2004). *A simulation study of the efficacy of inferential norming compared to traditional norming* (Assessment Report). San Antonio, TX: Harcourt.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Willingham, W. W. (1999). A systematic view of test fairness. In S. J. Messick (Ed.), *Assessment in higher education: Issues of access, quality, student development, and public policy* (pp. 213–242). Mahwah, NJ: Erlbaum.

Wood, J. M., Nezworski, M. T. & Stejskal, W. J. (1996). The Comprehensive System for the Rorschach: A critical examination. *Psychological Science, 7,* 3–10.

Woodcock, R. W. (1999). What can Rasch-based scores convey about a person's test performance? In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 105–127). Mahwah, NJ: Erlbaum.

Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 65–104). Mahwah, NJ: Erlbaum.

Wright, B. D. (2001). Separation, reliability, and skewed distributions. *Rasch Measurement Transactions, 14,* 786.

Wu, M. (2010). Measurement, sampling, and equating errors in large-scale assessments. *Educational Measurement: Issues and Practice, 29*(4), 15–27.

Zhu, J., Cayton, T., Weiss, L., & Gabel, A. (2008). *WISC-IV extended norms (WISC-IV technical report #7).* San Antonio, TX: Pearson Education. Retrieved from www.pearsonassessments.com/NR/rdonlyres/C1C19227-BC79–46D9-B43C-8E4A114F7E1F/0/WISCIV_TechReport_7.pdf

Zhu, J., & Chen, H. (2011). Utility of inferential norming with smaller sample sizes. *Journal of Psychological Assessment, 29*(6), 570–580.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Hillsdale, NJ: Erlbaum.