2

PSYCHOLOGICAL DIAGNOSTIC TESTING: ADDRESSING CHALLENGES IN CLINICAL APPLICATIONS OF TESTING

JANET F. CARLSON AND KURT F. GEISINGER

Psychologists have long been recognized as experts in developing, standardizing, and validating tests on which a variety of assessments rest. Individual achievement, ability, knowledge, personality, skills, and intelligence are only some of the dimensions that psychologists have worked to operationalize and measure reliably and accurately. Indeed, "psychological assessment has been a defining practice of professional psychology since the field's inception" (Camara, Nathan, & Puente, 2000, p. 141).

Although *tests* and *assessments* are not the same thing, a substantial portion of assessment activities depend on the use and interpretation of results obtained from individual tests. Assessment refers to the integration of information collected from a large number of sources, some of which likely include formal test data (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999). Tests often provide information that is essential to a comprehensive assessment. Good tests, used properly,

demonstrate that they are dependable and valid, a claim that cannot be made by less rigorously developed assessment processes, such as some observational and interview techniques.

Discussion within this chapter is limited to psychological diagnostic (or clinical) testing. For our purposes, we view this form of testing as comprising a major component of assessment processes that are applied within clinical contexts. Most of the contexts we consider involve treatment-based practices, such as mental health establishments and forensic facilities. For the most part, tests used in these settings would not be regarded as high stakes, although there may be exceptions on a case-by-case basis. Numerous assessment practices in educational contexts, with their attendant myths and fallacies, are addressed in the following chapters. Tests commonly associated with educational classification and eligibility determinations, such as placement tests or those used primarily to determine educational needs (e.g., to determine eligibility to receive special educational services) are not part of the discussion that follows. Our discussion centers on tests that are best described as clinical, personality, or psychodiagnostic in nature. Although we have not intentionally addressed either high-stakes educational testing or intelligence testing, some of the points we make may have relevance in educational and clinical neuropsychological arenas as well. Correspondingly, some points enumerated in the final chapter concerning large-scale cognitive diagnostic testing may resonate with portions of our discussion.

The kinds of measures we have envisioned while articulating our points are those that are administered as part of an assessment battery, given to an individual. The goals of such an assessment are to specify a diagnostic category or impression of some kind that will be used to formulate a treatment or intervention aimed at helping the test taker achieve a more satisfying level of functioning. Because prognosis is routinely considered a part of diagnostic efforts, prediction is important to our discussion. Of course, predictions often are influenced by many factors besides diagnoses. For example, the number of months a soon-to-be-discharged patient remains in the community may be affected directly by funding cuts to social programs, regardless of his or her diagnostic test results and regardless of his or her diagnosis at the time of discharge.

On an individual basis, psychological diagnostic testing involves a customized process in which several psychological measures are administered to individuals who are considered by themselves or by others to be in need of assistance or treatment for mental health issues. Measures may include a variety of techniques that produce useful and informative data—everything from interviews to objective tests that use optically scanned formats. As noted by Regier et al. (1998), "No assessment of clinical status is independent of the reliability and validity of the methods used to determine the presence of a diagnosis—be it by an unstructured clinical interview, a structured clinical assessment, or a highly structured instrument" (p. 114). To

ensure that conclusions suggested by the data are the result of a multivariate assessment paradigm, test users are obligated to use multiple data sources, which often include a clinical interview together with a number of tests selected on the basis of the test taker's presenting problems, characteristics, and needs (Garb, 2005; Yates & Taub, 2003). Unfortunately, although standards for testing require multivariate assessment (i.e., a "test battery"), "there is no empirical measure of the validity of a battery of tests . . . [in fact,] there is no uniform definition of the validity of a battery of tests" (Cates, 1999, p. 636).

Until recently, clinicians largely felt free to choose whichever tests they believed would serve their assessment purposes. Many practitioners who received training on specific measures during their graduate educations tended to use these same measures with their patients or clients (Cates, 1999). As a result, test batteries often were static, and some may have included tests that had little to do with a particular diagnostic quandary. Alternatively, some test batteries may have failed to include tests that would have been of considerable assistance in reaching a differential diagnosis. Evidence has accumulated since about the mid-1990s to indicate that assessment batteries have become less a force of habit. Today clinicians are exhorted to base test selection on the reason for the assessment and are more likely to use tests that address specific referral questions (Cates, 1999; Fong, 1995; Griffith, 1997; Meyer et al., 2001; Yates & Taub, 2003). This positive shift appears to be attributable in part to the constraints of managed care and third-party payers who require sound justifications for expenses submitted for reimbursement or prior approval.

Immediate goals of psychological diagnostic testing may include one or more of the following:

- to address more fully or accurately individuals' mental health needs,
- to improve treatment effectiveness,
- to guide interventions,
- to track treatment progress,
- to satisfy insurance requirements, and
- to satisfy managed care restrictions.

We believe that test results may be used effectively to achieve these aims. Psychodiagnostic tests can be used to promote accurate conceptualizations (diagnoses) of individuals who may be experiencing emotional discomfort or mental disturbance, whose ability to manage important aspects of living has deteriorated, whose ability to function in social or occupational spheres is impaired, or who may be demonstrating abnormal or bizarre behaviors or behaviors that are deeply disturbing or disruptive to others. Furthermore, we believe that psychological diagnostic testing may be used to help illuminate appropriate interventions for such individuals.

HISTORY OF CONTROVERSIAL ISSUES

Controversies surrounding clinical assessment have been around for many decades. One of the earliest, most influential, and well-reasoned set of observations was made by Meehl in 1954. Meehl's seminal work compared the accuracy of clinical and statistical approaches to prediction that together with other works in a similar vein (Ben-Porath, 1997; Grove & Meehl, 1996; Meehl, 1986) have cast doubt on the value of clinical judgment and expertise, particularly in light of equal or superior outcomes that could be produced using mechanical or actuarial approaches. Meehl, however, claimed that he did not intend to bring about the demise of clinical practitioners. In the preface to his reprinted work (Meehl, 1954/1996) and elsewhere (Meehl, 1986), he expressed what appears to be surprise and, perhaps, frustration with the various interpretations and simplifications of his original treatise, in which he went to considerable effort to present balanced evidence of merit on both "sides" of the argument. Meehl (1954/1996) stated that he "developed a certain Buddhist detachment" (p. xi) about the debate and seemed to take umbrage at suggestions that he was "grinding one axe or another" (p. x). Since the genesis of the clinical versus statistical debate, many writers—including Meehl (1986, 1954/1996)—have urged psychologists to abandon what developed into something of a red herring, but many elements of the debate linger and continue to manifest in various forms (Grove, Zald, Lebow, Snitz, & Nelson, 2000; Holt, 1986; Kleinmuntz, 1990; Meehl, 1986; Westen & Weinberger, 2004).

Intentional or not, the challenge posed by Meehl's early assertions helped to create an atmosphere that fostered closer scrutiny of clinical training, especially concerning diagnostic and assessment practices and judgments emanating from them (e.g., Cates, 1999; Garb, 1989). Nearly 40 years after the publication of Meehl's original work, Masling (1992, p. 53) observed that

> the ability to use psychological assessment methods is a unique and valuable skill in clinical psychology. Beyond that there is considerable controversy. It is something psychologists are uniquely qualified to do, but what it is intended to do and how well we do it remain unspecified.

Assessments involve integration of information from many sources: clinical interviews, behavioral observations, clinical histories, and, of course, psychological tests. Most clinicians depend on tests as an integral component of their assessment practices, although much evidence has accrued that substantiates a marked decline in test usage beginning near the latter part of the 20th century (Ben-Porath, 1997; Eisman et al., 2000; Meyer et al., 2001) and a corresponding decline in training in testing and assessment within graduate programs (Aiken, West, Sechrest, & Reno, 1990; Hayes, Nelson, & Jarrett, 1987). In response to a recent survey concerning test usage (Camara et al., 2000), clinical psychologists reported that they most frequently tested for

personality or diagnostic reasons (i.e., determination of presence or type of psychopathology). These findings are consistent with an earlier survey, cited by Camara et al. (2000), indicating that psychologists "primarily used assessments for diagnostic purposes, and 53% also used testing as an indicator of what type of therapy would be most effective" (O'Roark & Exner, 1989, as quoted in Camara et al., 2000, p. 142).

Few tests of personality or psychopathology rise to the level of being truly diagnostic instruments, such that a particular test result can be used to pinpoint a specific Axis I or Axis II disorder within the taxonomy of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*; American Psychiatric Association, 2000). Rather, many measures serve primarily as screening measures for particular symptoms or groups of symptoms that comprise essential features of certain mental disorders. Measures do not exist for all psychological diagnoses, of course, nor for every symptom or symptom cluster that may be indicative of the presence of particular disorders. It is unlikely that additional tests will be developed to such an extent that all—or nearly all—diagnoses are addressed, given the extensiveness of the current and recent *DSM* nosology. In addition, numerous codes and qualifiers exist within the *DSM* to indicate an unclear or incomplete symptom picture. For example, disorders may be designated as "atypical," "NOS" (not otherwise specified), or "unspecified mental disorder." When available information is not sufficient to render a differential diagnosis, the notation "diagnosis deferred" is made. Alternatively, the notation of "provisional diagnosis" is made when the clinician expects the individual will meet the specified criteria but does not do so currently, either because a full symptom picture cannot be provided by the client or because the duration of one or more symptoms falls short of that required by *DSM* criteria (and, with time, that requirement is expected to be met). Obviously, measures of any kind will be of little or no help in codifying these poorly differentiated, atypical cases. Likewise, conditions that may be the focus of treatment but that are not considered mental disorders (e.g., various relational problems, problems related to abuse or neglect) are coded as "V-codes" per *DSM* procedures. Such situations are not now, nor will they likely ever be, amenable to psychodiagnostic testing per se.

Measures used to screen for or identify particular *DSM* disorders often comprise items that closely mirror *DSM* criteria. Content validity of such tests generally rests on the extent to which the test scores and resultant interpretations align with essential diagnostic features. For example, the Beck Depression Inventory—II (BDI–II; Beck, Steer, & Brown, 1996), a screening measure for depressive symptomatology, represents a modification of an earlier version of this instrument. In the latter version, test directions and item content were revamped specifically to make them more consistent with diagnostic elements of the *DSM*.

It follows from the foregoing discussion that as diagnostic criteria undergo revisions, measures that had been developed to reflect those criteria

may not perform at the same level. By extension, truly diagnostic tests become less valid for diagnostic use as a function of revisions to diagnostic criteria. Changes in diagnostic criteria may lead to dramatic differences in reported prevalence rates (Regier et al., 1998) that have little to do with genuine epidemiological changes. Nevertheless, differential rates of diagnosis may lead to more or less research funding for particular disorders and assessment. Also, public attention may follow apparent upsurges in the incidence of particular psychological disorders. The legitimacy of classification systems has been questioned many times and has stirred up much controversy over the years. These issues call into question the value and probity of diagnoses as well as the precise definitions of mental illness and mental health. Although these issues are beyond the scope of this chapter, they comprise the backdrop against which the discussion occurs and also create ongoing pressure for revisions to the current taxonomy of mental disorders. Psychological diagnostic tests will need to keep apace with revisions prompted by this discourse.

The remainder of this chapter addresses specific myths and fallacies associated with psychological diagnostic testing, disregarding as much as possible the controversies surrounding clinical diagnosis. When controversies have relevance to a particular point of discussion, we draw on the controversy to enrich and frame our position rather than try to resolve long-standing debates, many of which have no actual answers. We have imagined that the controversies, myths, and fallacies described could be proffered by a variety of individuals or groups, at least some of whom may use tests or may be considering using them.

CURRENT CONTROVERSIES, MYTHS, AND FALLACIES

Several challenges have emerged since the 1990s concerning clinical assessment and psychodiagnostic testing. A number of these issues can be traced to the expanding influence of managed mental health care and third-party payers. These ubiquitous enterprises operate from a business framework, with emphasis placed on cost containment, cost–benefit analyses, and accountability. Clearly, third-party reimbursement and the stipulations of managed care affect the use of psychological tests for the purpose of clarifying, rendering, or differentiating clinical diagnoses. It is understandable that third parties are disinclined to pay for services solely on the basis of treating psychologists' requests. Clinicians must propose and justify to payers interventions that follow logically from the particulars of the case under treatment. Insurers have treated requests by practitioners for reimbursement of services involved in the administration of diagnostic tests in the same way as more traditional therapeutic interventions. As far as practitioners are (or should be) concerned, a number of researchers have demonstrated the therapeutic

value of clinical tests and have encouraged the use of such tests as interventions in and of themselves (Ben-Porath, 1997; Finn & Tonsager, 1997; Hayes et al., 1987).

Substantial evidence supports the treatment utility of assessments, making the "notable decrease in the clinical use of psychological testing" (Finn & Martin, 1997, p. 374) observed by several authors (e.g., Camara et al., 2000; Finn & Martin, 1997; Griffith, 1997; Kubiszyn et al., 2000) perplexing and difficult to explain. Some writers have suggested that the lack of effective public relations within the helping professions is partly to blame; for example, Finn and Martin (1997) suggested that a "confusing message [has been sent] to non-psychologists about the value of psychological assessment" (p. 374).

Among the fallacious beliefs and misconceptions about psychological diagnostic tests are the ideas that (a) tests are too costly to justify their use, (b) tests are not valid, (c) tests cannot be used effectively in multicultural contexts, (d) test content is peculiar and irrelevant, and (e) tests are easily manipulated by respondents who wish to engineer a particular outcome. Each of these myths is addressed in the following discussion.

Psychodiagnostic Tests Are Too Expensive

Costs of psychodiagnostic testing can be substantial for several reasons, many related to test development issues. Test development is a costly endeavor, especially if one adheres faithfully to existing standards. Many reputable test publishers do attend carefully to these standards, and it is not surprising that costs associated with doing so are largely passed along to the consumer, as with other products on the market. The standards for testing (AERA, APA, & NCME, 1999) establish policies to guide the development of tests in ways that ensure their relevance, dependability of scores, and validity of their use for specific purposes.

A well-developed test depends on input from numerous experts, such as content or measurement specialists, some who write items for possible inclusion on the test under development and others who review items for potential bias and the extent to which the content domain has been sampled adequately. After a pool of items has been developed, item tryouts are conducted, and some items are culled because of their poor empirical performance. Other items are revised or edited in some way. For norm-referenced tests, the next step involves obtaining a group of individuals on whom test norms are developed. This group must be similar to the group for whom the test is intended for use. Often this involves a national (or international) sample, with demographic characteristics that correspond with appropriate U.S. Census Bureau data. Tests that are well grounded psychometrically achieve this status by amassing evidence that supports their reliability and validity. Most often, this evidence derives from empirical studies that exam-

ine features such as test–retest and interscorer reliability, internal consistency estimates, and construct and other forms of validity, such as the test's ability to classify correctly a recognized clinical group. For instance, the BDI–II, described earlier, would be evaluated for its ability to correctly differentiate depressed from nondepressed individuals. Many psychodiagnostic measures contain subscales that represent components of the larger construct. For example, as a measure of personality, the Minnesota Multiphasic Personality Inventory—2 (MMPI–2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989) consists of 10 basic clinical scales that represent specific personality aspects, each of which may be developed to greater or lesser extents within a particular individual. The structure of such scales is verified using sophisticated factor-analytic techniques.

The foregoing description is offered to provide an overview of the processes involved in the development of measures, such as those developed for psychodiagnostic purposes. It represents a thumbnail sketch of the major steps that a relatively straightforward diagnostic test would undergo as it is being shaped into a psychometrically sound and marketable measure of personality or some aspect of personality. Essentially, we agree that well-developed tests are expensive, but we do not necessarily believe that the costs are excessive in light of all that goes into their development, not to mention the prospective value of information gleaned from the test results. Ultimately, the use of such measures may reduce costs because they quickly and efficiently gather information, thereby expediting treatment that is targeted appropriately for the problem.

Some readers may believe that clinicians decide to "run some tests" more as a way to pad their bills than to obtain essential information about a client with whom they are working. Not only would such behavior constitute a serious abrogation of ethical principles of practice (American Psychological Association, 2002), it would also be fiscally unsound as well. Results from a survey conducted by Camara et al. (2000) indicate that experienced clinical psychologists take more than 3.5 hours to administer, score, and interpret a full psychological battery. Assessment services are not always reimbursable by insurance. When these services are reimbursed, survey respondents indicated that reimbursement limits of less than 2 hours are most typical, leaving an average of 1.5 hours of assessment services not reimbursed for each individual for whom a full assessment battery is completed.

During the late 1980s and early 1990s, computer-based test administration, scoring, interpretation, and report-generation software burgeoned and prompted some writers to opine that more clinicians would avail themselves of these options to counteract diminishing insurance reimbursements. However, data from Camara et al. (2000) reveal that most clinicians continue to administer, score, and interpret tests themselves. Few respondents reported turning over administration or scoring responsibilities to someone with less training, despite some earlier concerns that this practice would become com-

monplace. Similarly, few reported that they invoked computer-based interpretation or report-generation programs.

At least in terms of time needed to conduct a comprehensive assessment versus time compensated, the balance does not favor those who administer tests. The decline in the use of tests may be related to a fiscal reality concerning a less than favorable ratio of time spent to hours reimbursed (e.g., Camara et al., 2000; Garb, 2003). Griffith (1997) examined this issue by interviewing several of the major third-party payers with regard to their practices and offered guidance for ways in which psychologists should justify interventions, including assessment activities. She stressed the need to speak the language of the payers by relating costs to benefits. In a similar vein, other authors have noted that it is essential to emphasize the effectiveness of assessment and to present the rationale in objective, financial terms that demonstrate the utility of assessments in improving treatment (Yates & Taub, 2003).

It is important to recognize that asking questions of an individual test taker represents a form of data collection for which a number of standardized instruments exist. Data from structured, semistructured, or unstructured interviews are frequently part of the assessment process. Many questions are posed directly to the individual being assessed, and his or her responses comprise an important component of many assessment batteries. Some may question the value of tests, above and beyond information that could be gleaned as easily by direct inquiry of the test taker. This question also revolves around the issue of validity, but not in the same vein as has been discussed thus far. Those who challenge the "value added" by data derived from a standardized test do not necessarily challenge whether the test achieves its intended purpose. Rather, they take issue with whether there is a large enough gain to be made beyond what is readily available from sources other than tests, especially in light of cost considerations. Arguments of this kind hinge on a cost–benefit analysis and have been used to impugn the incremental validity of tests (Finn & Martin, 1997; Garb, 2003; Hunsley & Meyer, 2003; Yates & Taub, 2003).

Garb (2003) discussed incremental validity with respect to the assessment of psychopathology using a variety of data collection procedures, including interviews, personality inventories, projective techniques, and brief self- and clinician-rated measures. He found that diagnostic efforts were aided by the use of self-report inventories in addition to conducting clinical interviews. Garb noted as well that projective techniques such as the Rorschach were helpful in more circumscribed areas, such as evaluating the presence of a thought disorder. In terms of personality and psychopathology, incremental gains were evident when a variety of assessment information was collected, including information from interview data, personality inventories, projective techniques, and self- and clinician-rated measures. Using a multifaceted assessment approach provides the most complete understanding of the clinical symptom picture. Whether critics choose to be influenced by

statistical findings of significance in this regard amounts to an idiosyncratic decision. Although treatment outcomes should be maximally and positively affected by assessments undertaken to establish a favorable cost–benefit ratio, many clinical outcomes are not monetary per se and "cannot be transformed into monetary units" (Yates & Taub, 2003, p. 480). Yates and Taub (2003) urged the inclusion of such nonmonetary factors in considering costs and benefits in stating that "the desired outcomes of many assessments are . . . reliable and valid descriptions of the status of psychological, biological, and social variables that are of potential use in delivering clinical services" (p. 480).

In assessing incremental validity, several authors have argued (or implied) that it is important to bear in mind the real world within which clinicians operate. Specifically, Garb (1989) described some ways in which studies of clinical versus mechanical prediction typically do not mirror clinicians' everyday experiences. In the studies reviewed by Garb, "judges were given a set of protocols and instructed to perform a specific task. In actual clinical practice, however, no one tells clinicians what judgments need to be made or what information they should obtain" (p. 392). In further discussion, the author noted that diagnostic accuracy was substantially higher for experts than for nonexperts when certainty of one's judgment was considered. This finding brings up a realistic issue—diagnostic certainty—that may be reflected in clinical practice to a much greater extent than can be captured in studies such as those reviewed by Garb. In practice, clinicians denote diagnostic uncertainty in several ways. For example, they may indicate that one or more other diagnoses need to be considered and ruled out before the differential diagnosis can be made. Alternatively, they may note that diagnosis is "deferred." Both differential and deferred diagnoses are coded (that is, made part of the clinical record) and remain in force, often pending receipt of additional information that may involve a physical examination, comprehensive medical history, consultation with family to corroborate certain facts, and so forth. Furthermore, clinicians in practice should not render diagnoses until the data are sufficiently clear. In research such as that described by Garb (1989), clinicians offered diagnoses on the basis of test protocols and instructions handed to them and could not seek out information they would need "in real life" to make a diagnosis.

Psychodiagnostic Tests Are Not Valid

The notion of validity of course requires a system of prediction and of understanding those predictions in the context of a model. At least four fallacious criticisms of personality tests emerge under the general theme of validity. The first of these relates primarily to the sheer number of personality measures and theories. The second relates to the match between personality theories and measures. The third relates to a general lack of consensus about

the criteria against which to evaluate personality tests and whether personality test data should be combined using clinical or statistical means. A fourth is discussed more fully in the section following this one: the apparent generalizability of personality approaches and tests across languages and cultures. Each of the other three is addressed briefly here.

The plethora of different models of personality and of personality assessment have led to a sense in the profession that when there are so many models, none can be truly valid or representative of the human person. In fact, all have some semblances of truth.

In his classic book on personality measurement, Wiggins (1973) concluded,

> An important but as yet unanswered question concerns the extent to which personality theories may facilitate the prediction of human behavior in applied settings. . . . In order for a personality theory to facilitate prediction, it must be relevant to the particular criterion situation, must be explicit enough to suggest the selection of specific testing instruments, must have definite implications for criterion performance, and must be capable of being evaluated. . . . [A personality test] can be justified only if it can be shown to produce incremental validity and a greater generality of application than does the use of procedures not guided by theoretical considerations. (p. 513)

Wiggins (1973) continued by describing a continuum of approaches to personality assessment and prediction. He identified the analytic approach as the one that places the greatest reliance on the theories of personality that undergird the assessments and their results. At the opposite extreme is the empirical approach, sometimes called *dustbowl empiricism*. This approach places the least weight on a theoretical approach to personality and focuses simply on the identification of different criterion groups or leads to the successful predictions of important criteria. Some criticisms of psychological tests have occurred from either end of the spectrum. Those holding dear the analytic approach find the empirical approaches lacking in explanation or theory; those embracing the empirical model simply seem not to understand the need for theory when their methods appear to work so much better. In an era when construct validity is preeminent, both theory (explanation) and predictions are required (Geisinger, 1992).

Regardless of the model, science requires both predicting and understanding the relationships among variables or constructs. As Wiggins (1973) concluded,

> The decision to rely on explicit theoretical considerations at various stages of the basic prediction paradigm does not commit the assessment psychologist to any particular theoretical model of personality. Nevertheless, despite predilections of any one psychologist for one type of theorizing over others, the personality models that are likely to be of greatest value to him [*sic*] are those whose principal constructs are most directly translatable into concrete testing procedures. (p. 514)

It is true that there is not a consensus about the criteria against which to evaluate personality tests. There are several reasons for this lack. One is that personality measures have been used in making a wide range of predictions, from psychodiagnosis to hiring firefighters and police officers, from making decisions regarding the nature of one's incarceration to evaluating one's likelihood of succeeding in college. Needless to say, each of these purposes requires a different kind of criterion. Even within single instances of use—say, in helping to decide whether to accept a college applicant (e.g., Sedlacek, 2004; Willingham, 1985)—success can be measured in numerous ways, from the traditional grade-point average, to completion of the college degree, to one's level of participation in nonacademic activities while in college. This fallacy needs to be understood rather as the perception that the tests have been seen as so robust that they have the potential for a wide range of predictions. Indeed, they have been found lacking in some settings. Even valid and appropriate medical tests cannot be used indiscriminately for a wide range of diseases. Using Wiggins's caution cited earlier, however, individuals should only use personality tests when theory suggests that the constructs are influential in leading to success.

Meehl's (1954/1996) work has been cited extensively to indicate that clinicians likely are not able to predict certain outcomes as well as a regression equation when clinicians and computers are provided the same information. An analogy to this finding is that even professional racetrack drivers cannot determine the velocity of a speeding car as well as radar. However, the driver is able to sense danger, whether the car is approaching its maximum performance and limits, and so on. Meehl did not suggest that clinicians should not make such predictions, only that they rely on statistical procedures and evidence whenever possible:

> Nobody disputes that it is possible to improve clinicians' practices by informing them of their track records actuarially. Nobody has ever disputed that the actuary would be well advised to listen to clinicians in setting up the set of variables. (Meehl, 1986, p. 372)

Thus Meehl certainly affirmed the use of personality measures in psychodiagnostic assessment. In fact, he ratified it and saw an important role for assessment psychologists as well. Moreover, he strongly affirmed the importance of clinicians engaging in therapy rather than making rare kinds of predictions (e.g., predictions to a parole board regarding the likelihood of recidivism).

Psychodiagnostic Tests Lack Cultural Generalizability

One myth of personality tests relates to the extent to which the results from such measures are culturally bound. One position holds that all measures are limited culturally to the country or culture in which they were built,

validated, and used extensively. Indeed, the translation, or better, adaptation of measures has been a topic of great interest in recent years as evidenced by a conference held at Georgetown University in 1999; a book edited by Hambleton, Spielberger, and Merenda (2005) based primarily on the papers presented at that conference; the biannual conferences of the International Test Commission; and the guidelines that that commission has developed guiding the translation and adaptation of psychological and educational measures (see Coyne, 2001). The term *adaptation* is preferred to *translation* because changing a test from one language, one country, and one culture to another involves more than a linguistic translation of the measure. It typically also involves extensive knowledge of the culture, the characteristics being assessed and how they are manifested in both cultures, and subtleties of their interaction.

The guidelines are organized into several sections, relating to context (2 guidelines), test development and test adaptation (10 guidelines), test administration (6 guidelines), and documentation and test score interpretations (4 guidelines). The guidelines caution users against blindly translating and using tests from one language and culture to another. For instance, one must take into account cultural and semantic differences in the measures. Two test development and adaptation guidelines state, for example, "Test developers/publishers should provide evidence that the language use[d] in the directions, rubrics, and items themselves as well as in the handbook are appropriate for all cultural and language populations for whom the instrument is intended" (Coyne, 2001, ¶ D.2), and "Test developers/publishers should provide statistical evidence of the equivalence of questions for all intended populations" (Coyne, 2001, ¶ D.9). Similarly, a documentation and score interpretation guideline suggests that "when a test is adapted for use in another population, documentation of the changes should be provided, along with evidence of the equivalence" (Coyne, 2001, ¶ I.1). Thus these guidelines suggest that one must provide evidence that scales can be taken from one culture and used in another. Such efforts typically involve the translation of the test but also involve cultural adaptation. For ease of explanation, an example from an intelligence test for children is provided. Imagine an item from such a scale that asks a child to explain a proverb such as, "A stitch in time saves nine." Merely changing the language from the original or source language to a target language would not make the item equivalent to the original, and as part of such a translation, its level of difficulty and underlying meaning would be radically changed. Evidence is needed that the results of testing are equally meaningful in the new language and culture. Such evidence is far more comprehensive than simply validating a test. Rather, one must show that the measure is valid (e.g., in a predictive sense) in the new language and culture and also that the scores are similarly or equivalently meaningful in the new language. In other words, one must show that the validity of the use of the inventory or test in the new culture generalizes

to or parallels that of the original culture. More complete descriptions of some of the issues involved in adapting tests across languages and cultures and in interpreting the results of such test use may be found in Geisinger (1994, 1998).

Because of the desire for efficiency among test publishers and researchers, evidence supportive of the use of some measures in different languages and cultures is currently emerging. Butcher, Cabiya, Lucio, and Garrido (2007; see also Butcher, 2004) have provided substantial evidence that the MMPI–2, one of the most frequently used measures in psychological settings in the United States (Camara et al., 2000), has highly similar validity and uses for Americans in the United States who come from Latin cultures. This evidence is based on research with such adaptations among Spanish-speaking clients in both the United States and Latin America. Moreover, such evidence is not limited to a single Spanish translation but to a series of them, often specific to different countries in Latin America. Discussions that psychopathology may be less culture specific than many have believed can be found in Butcher, Coelho Mosch, Tsai, and Nezami (2006) and Good and Kleinman (1985); these researchers reported that psychiatric diagnoses are likely not culture specific but that there may be differing manifestations in symptomatology between cultures.

Although literally hundreds of personality measures have been developed in the past century, personality theorists appear to have come to considerable consensus that a five-factor theory of personality subsumes most theories (e.g., Goldberg, 1990, 1993). The primary proponents of this model, McCrae and Costa (1999), although not contending that the five factors (Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience) are the only aspects of personality, have fostered research to demonstrate that it appears these factors have generalizability across cultures (e.g., McCrae, 2000, 2001; Yang et al., 1999). The emergence of the five-factor theory has driven much research on personality assessment at least since the 1990s, with many measures—even some that were developed using very different conceptualizations of personality—having shown a degree of correspondence with this model.[1]

In summary, it appears that at least some psychological and personal constructs do appear relevant across cultures. Moreover, measures developed in one culture and country can be adapted to other cultures. We know that at least some of these measures are able to be used with success cross-culturally. One does still need, however, to follow the guidelines of the International Test Commission (see Coyne, 2001). In so doing, one must translate into the target language while keeping the measure sensitive to the target culture

[1]Some Eastern cultures have found a sixth factor in addition to these five. This sixth factor, which has been found in Chinese adaptations of personality instruments and has been seen in other Asian populations as well, has been identified as Interpersonal Relatedness (Cheung et al., 2001).

associated with that language and carefully document any changes made. Evidence demonstrating the equivalence of the instrument in the new culture is required. For some successful measures, however, it appears highly likely they are able to be used in a good number of cultures. Merely demonstrating that a measure is useful in several cross-cultural settings does not prove that it can be used in all cultures, but it does provide a hypothesis that it may be worth the effort to adapt.

Psychodiagnostic Test Content Is Peculiar and Irrelevant

Most of the discussion so far has focused on tests as likely elements of a larger assessment quandary pertaining to an individual. Evidence collected from standardized testing may be used to corroborate evidence collected from other sources. Attempting to evaluate the relevance of test content by removing the test from the context within which it is normally embedded may distort and underestimate its value. Test results must be integrated to form a coherent diagnostic picture. This point was made artfully by Cates (1999), who stated,

> The accumulation of data is not an assessment. The integration of data and subsequent interpretations comprise the assessment. The alphabetically arranged list of words, "and," "go," "I," "let," "then," "us," and "you" mean little. Arranged as "Let us go then, you and I . . . " they form the introductory line to one of the most famous poems of our century. (p. 638)

An analogous situation for our purposes might be seen if one were to assess treatment progress for an individual with an anxiety disorder purely on the basis of a score on a single measure of anxiety that indicates the presence or absence of anxiety-related symptoms. To understand fully the degree of improvement or decline for a specific patient, one would need to compare current levels of anxiety with both previous levels and other outcome measures, such as symptom duration and the extent to which symptoms interfere with other aspects of the client's life, such as the ability to reenter the workforce and maintain employment. However, because psychometric procedures are not available for multiple tests taken together, the psychometric soundness of individual tests continues to be evaluated on a test-by-test basis (and for specific populations and subpopulations) rather than as one component of the larger assessment process. Thus, in the anxiety example, it is highly likely that treatment progress will be gauged solely in terms of the presence or absence of anxiety-related symptoms, such as may be measured by a single inventory. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) require a multivariate approach to assessment, but it is the individual tests that must each pass muster rather than the group of measures that actually might be used in a given assessment battery.

A related problem may occur when a test taker (or parents of a test taker, in the case of school-aged children) single out an item or subtest and question its inclusion. Frequently, such individuals inquire about the relevance of particular items or subtests for the characteristic being assessed. In essence, queries that arise in this way challenge a test's face validity but do so in a somewhat unorthodox—and possibly biased—way. It is likely that the questions recalled in these instances are the ones that stand out in the individual's recollection, because they were perceived to be unusual or unexpected in light of the test taker's understanding of the test's purpose. Typically, test takers do not recall and question items that appear to belong on the test—that is, are *face valid*. Face validity offers weak support for a measure, so much so that it is probably inappropriate to discuss it as a psychometric property. Face validity may do little to enhance the technical features of a given test, although higher levels of face validity may promote greater effort on the part of test takers if their belief in the test motivates them or makes them try harder. A substantial concern arises when items are viewed in isolation from the test itself. The functioning of individual items is important during test development, but the value of individual items is greatly diminished after the entire test has been built and validated. A comparable situation could occur when a physician records a person's height and that patient questions the relevance of his or her height as far as the determination of health status. Of course, when height is considered together with weight (and age and sex), a more complete assessment is achieved, which can serve as an important determinant of physical health status that also may affect estimates of risk for developing various medical conditions. A published test is not intended to be used in fragmented bits and pieces, and findings obtained from any sort of nonstandardized administration cannot rightly be considered accurate (Geisinger & Carlson, 2002). Test takers who question the content of tests should be advised of the reality that the entire test forms the basis for conclusions about his or her personality or psychological state of being.

Specialized scales developed for or from published tests (such as the myriad supplemental scales of the MMPI that assess for dominance, addiction potential, hostility, social responsibility, marital distress, and many other aspects of personality) are another matter entirely. Often these scales were developed in applied clinical settings, with the intention that they would have direct clinical relevance for patients and clients in those or highly similar settings. Many of these scales have undergone rigorous test development and psychometric scrutiny comparable to the originally published test. For example, Keane, Malloy, and Fairbank (1984) developed the Posttraumatic-Stress Disorder (PK) scale to assess for the presence of posttraumatic stress disorder (PTSD) using MMPI item responses of 60 male Vietnam combat veterans previously diagnosed with PTSD. Their responses were contrasted with a similar group of veterans who were diagnosed with disorders other

than PTSD. The two groups answered 49 items differently, and these items formed the basis of the PK scale, which subsequently was evaluated for its ability to correctly classify veterans with and without PTSD and for its psychometric properties of reliability and validity.

Psychodiagnostic Tests May Be (Easily) Faked

Questions about test accuracy extend beyond questions about tests themselves to include the test setting and the test taker. The possibility that test takers may respond in less than truthful ways undercuts the confidence some people may have in test results and interventions derived from these results. In situations in which test takers have little to gain by practicing deception while responding to test items, face-valid tests may enhance rapport and motivation on the part of the test taker. The straightforward nature of such test items may encourage honest responding. For example, a patient who complains at intake about a sullen or dysphoric mood, sleep disturbances, and a sense of foreboding about the future may feel that a psychologist who administers a depression inventory understands the symptom picture and is attempting to gain more information that will be helpful in formulating effective interventions. Under these circumstances, the patient is likely to respond in accordance with his or her perceptions.

Under other circumstances, however, test takers indeed may be motivated to present themselves in a particular manner that does not accurately reflect their personality or clinical condition. Some tests comprise items that lend themselves to easy manipulation by test takers, because the items make obvious the purpose of the test and the likely manner in which results will be used. Prior and subsequent to test administration, psychologists must consider whether there is a discernible reason for test takers to be less than forthright in their responses and whether that reason might constitute a motive for faking. If so, the test giver must choose tests and interpret test findings with these possibilities in mind. In child custody debates, for example, one or both parents may be assessed for parental fitness. Either or both of them may attempt to present themselves in a too-favorable light. In an extreme case, either may attempt to exaggerate his or her stability in an effort to win custody. In such cases, using a test that is patently obvious in its intent would be inadvisable, because the motivation to "fake good" is apparent. Situations that may prompt test takers to "fake bad" are imagined easily within the context of the legal system or within forensic settings. An individual accused of a heinous crime may prefer to be regarded as "insane" instead of "guilty" and could attempt to distort his or her test responses accordingly. Malingering detection in such cases is crucial to protect the rights of all parties involved—victims and victims' families, as well as individuals who truly suffer from mental disorders such that they are unable to appreciate the wrongfulness of their acts or are unable to conform their behavior to legal standards.

Many self-report inventories have addressed the issue of faking by employing deception detection methods, which may take the form of validity scales embedded within the tests themselves. The first test to use this approach was the original MMPI (Hathaway & McKinley, 1943). Hathaway and McKinley added 4 validity scales to the 10 clinical scales to ensure that various test-taking attitudes (such as the tendency to present oneself in an overly positive manner) were understood and could be brought to bear in the interpretive process. It is interesting that the test authors built the MMPI using an empirical keying method of test construction, which greatly reduces the likelihood that such distortion will occur. This method let the data speak for themselves by using statistical analyses to determine scales to which an item contributes rather than using clinical judgment. Thus it is difficult for test takers to know what the test items are assessing. Even so, some authors have urged caution, especially when subsets of items from larger inventories are administered separately rather than by being culled from an administration of the entire inventory. The PK supplementary scale of the MMPI cited earlier, for example, may be susceptible to faking by veterans who wish to receive benefits or compensation (Fairbank, McCaffrey, & Keane, 1985). Since the publication of the original MMPI, many other personality measures have included validity scales to detect—and even correct for—response tendencies demonstrated by test takers.

CONCLUSION

As noted in this chapter, standardized assessment batteries are virtually nonexistent, even though the use of multiple assessment methods is routine and highly recommended by the *Standards* for testing (AERA, APA, & NCME, 1999). "The new challenges to this field will be to standardize assessment methods and to specify the scope of clinically significant disorders that are in need of treatment" (Regier et al., 1998, p. 110). It seems that much would be gained by the development of psychometrically sound assessment batteries. Certainly, diagnostic accuracy would be improved by greater standardization of assessment methods used for psychodiagnostic purposes.

> If some greater degree of standardizing assessment methods can . . . be accomplished, psychiatric epidemiological and services research will be positioned for another leap forward in both improving the understanding of the cause(s) of mental disorder and helping to focus limited resources in the most cost-effective manner. (Regier et al., 1998, p. 115)

Issues emanating from managed care, the age of accountability, and cost containment (Ben-Porath, 1997; Griffith, 1997; Meyer et al., 2001; Yates & Taub, 2003) are likely to be with us out into the foreseeable future. In this regard, treatment utility emerges as a key consideration because it provides

the kind of model that is responsive to the accountability and cost–benefit demands embodied by managed mental health care (Ben-Porath, 1997; Finn & Tonsager, 1997; Griffith, 1997). This model casts assessment itself as a therapeutic intervention. In 1997, Finn and Tonsager observed that "treatment utility of assessment is weaker than many of us might want" (p. 375). Since then, evidence has continued to surface that suggests that clear gains have been made in establishing the effectiveness of this approach (Nelson-Gray, 2003).

Assessment is a human endeavor and carries with it the promises and pitfalls of being so. Practitioners have been unwilling, or at least slow, to embrace mechanical prediction models, which themselves have been slow to materialize (e.g., Kleinmuntz, 1990). Although the idea of highly accurate (mechanical) predictions is alluring in some respects, it may also be disturbing to some who may find it difficult to imagine making clinical decisions by a purely formulaic "crunching of numbers."

Much of the foregoing material that has focused on points of contention with regard to clinical diagnostic testing has also served to advance the field and to illuminate future directions for clinical research and practice related to psychodiagnosis. Informed assessment practices that use the best available measures in combination with one another, that build on a collaborative assessment process inclusive of the patient or client and are mindful of his or her culture, that identify the most effective interventions for that individual, and that practitioners are reasonably compensated for is an ideal toward which to strive.

REFERENCES

Aiken, L. S., West, S. G., Sechrest, L., & Reno, R. R. (1990). Graduate training in statistics, methodology and measurement in psychology: A survey of PhD programs in North America. *American Psychologist, 45*, 721–734.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: Author.

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

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

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Beck Depression Inventory manual* (2nd ed.). San Antonio, TX: Psychological Corporation.

Ben-Porath, Y. S. (1997). Use of personality instruments in empirically guided treatment planning. *Psychological Assessment, 9*, 361–367.

Butcher, J. N. (2004). Personality assessment without borders: Adaptation of the MMPI–2 across cultures. *Journal of Personality Assessment, 83*, 90–104.

Butcher, J. N., Cabiya, J., Lucio, E., & Garrido, M. (2007). *Assessing Hispanic clients using the MMPI–2 and MMPI–A*. Washington, DC: American Psychological Association.

Butcher, J. N., Coelho Mosch, S., Tsai, J., & Nezami, E. (2006). Cross-cultural applications of the MMPI–2. In J. N. Butcher (Ed.), *MMPI–2: A practitioner's guide* (pp. 505–537). Washington, DC: American Psychological Association.

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

Camara, W. J., Nathan, J. S., & Puente, A. E. (2000). Psychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice, 31*, 141–154.

Cates, J. A. (1999). The art of assessment in psychology: Ethics, expertise, and validity. *Journal of Clinical Psychology, 55*, 631–641.

Cheung, F. M., Leung, K., Zhang, H. Z., Sun, F. A., Gan, Y. Q., Song, W. Z., & Zie, D. (2001). Indigenous Chinese personality constructs: Is the five-factor model complete? *Journal of Cross-Cultural Psychology, 32*, 407–433.

Coyne, I. (2001, April 21). *ITC test adaptation guidelines*. Retrieved July 1, 2008, from http://www.intestcom.org/test_adaptation.htm

Eisman, E. J. Dies, R., Finn, S. E., Eyde, L. D., Kay, G. G., Kubiszyn, T. W., et al. (2000). Problems and limitations in the use of psychological assessment in contemporary health care delivery. *Professional Psychology: Research and Practice, 31*, 131–140.

Fairbank, J., McCaffrey, R., & Keane, T. (1985). Psychometric detection of fabricated symptoms of post-traumatic stress disorder. *American Journal of Psychiatry, 142*, 501–503.

Finn, S. E., & Martin, H. (1997). Therapeutic assessment with the MMPI–2 in managed health care. In J. N. Butcher (Ed.), *Objective personality assessment in managed health care: A practitioner's guide* (pp. 131–152). New York: Oxford University Press.

Finn, S. E., & Tonsager, M. E. (1997). Information-gathering and therapeutic models of assessment: Complementary paradigms. *Psychological Assessment, 9*, 374–385.

Fong, M. L. (1995). Assessment and *DSM–IV* diagnosis of personality disorders: A primer for counselors. *Journal of Counseling & Development, 73*, 635–639.

Garb, H. N. (1989). Clinical judgment, clinical training, and professional experience. *Psychological Bulletin, 105*, 387–396.

Garb, H. N. (2003). Incremental validity and the assessment of psychopathology in adults. *Psychological Assessment, 15*, 508–520.

Garb, H. N. (2005). Clinical judgment and decision making. *Annual Review of Clinical Psychology, 1*, 67–89.

Geisinger, K. F. (1992). The metamorphosis in test validation. *Educational Psychologist, 27*, 197–222.

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6,* 304–312.

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.

Geisinger, K. F., & Carlson, J. F. (2002). Standards and standardization. In J. N. Butcher (Ed.), *Practical considerations in clinical personality assessment* (2nd ed., pp. 243–256). New York: Oxford University Press.

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology, 59,* 1216–1229.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48,* 26–34.

Good, B., & Kleinman, A. (1985). Epilogue: Culture and depression. In A. Kleinman & B. Good (Eds.), *Culture and depression* (pp. 491–506). Berkeley: University of California Press.

Griffith, L. (1997). Surviving no-frills mental health care: The future of psychological assessment. *Journal of Practical Psychiatry and Behavioral Health, 3,* 255–258.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law, 2,* 293–323.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12,* 19–30.

Hambleton, R. K., Spielberger, C. D., & Merenda, P. F. (2005). *Adapting educational and psychological tests for cross-cultural assessment.* Hillsdale, NJ: Erlbaum.

Hathaway, S. R., & McKinley, J. C. (1943). *The Minnesota Multiphasic Personality Inventory.* Minneapolis: University of Minnesota Press.

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

Holt, R. R. (1986). Clinical and statistical prediction: A retrospective and would-be integrative perspective. *Journal of Personality Assessment, 50,* 376–386.

Hunsley, J., & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological Assessment, 15,* 446–455.

Keane, T. M., Malloy, P. F., & Fairbank, J. A. (1984). Empirical development of an MMPI subscale for the assessment of combat-related post-traumatic stress disorder. *Journal of Consulting and Clinical Psychology, 52,* 888–891.

Kleinmuntz, B. (1990). Why we still use our heads instead of formulas: Toward an integrative approach. *Psychological Bulletin, 107,* 296–310.

Kubiszyn, T. W., Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., et al. (2000). Empirical support for psychological assessment in clinical health care settings. *Professional Psychology: Research and Practice, 31,* 119–130.

Masling, J. M. (1992). Assessment and the therapeutic narrative. *Journal of Training and Practice in Professional Psychology,* 6, 53–58.

McCrae, R. R. (2000). Beginning again on personality and culture [Review of the book *Personality and person perception across cultures*]. *Contemporary Psychology, 45,* 38–40.

McCrae, R. R. (2001). Trait psychology and culture: Exploring intercultural comparisons. *Journal of Personality,* 69, 819–846.

McCrae, R. R., & Costa, P. T., Jr. (1999). A five-factor theory of personality. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality psychology* (pp. 139–155). New York: Guilford Press.

Meehl, P. E. (1986). Causes and effects of my disturbing little book. *Journal of Personality Assessment, 50,* 370–375.

Meehl, P. E. (1996). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* Northvale, NJ: Jason Aronson. (Original work published 1954)

Meyer, G. J., Finn, S. E., Eyde, L., Kay, G. G., Moreland, K. L., Dies, R. R., et al. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist,* 56, 128–165.

Nelson-Gray, R. O. (2003). Treatment utility of psychological assessment. *Psychological Assessment, 15,* 521–531.

O'Roark, A. M., & Exner, J. E. (Eds.). (1989). *History and directory: Society for personality assessment fiftieth anniversary.* Hillsdale, NJ: Erlbaum.

Regier, D. A., Kaelber, C. T., Rae, D. S., Farmer, M. E., Knauper, B., Kessler, R. C., & Norquist, G. S. (1998). Limitations of diagnostic criteria and assessment instruments for mental disorders. *Archives of General Psychiatry,* 55, 109–115.

Sedlacek, W. E. (2004). *Beyond the big test: Noncognitive assessment in higher education.* San Francisco: Jossey-Bass.

Westen, D., & Weinberger, J. (2004). When clinical description becomes statistical prediction. *American Psychologist,* 59, 595–613.

Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment.* Reading, MA: Addison-Wesley.

Willingham, W. W. (1985). *Success in college: The role of personal qualities and academic ability.* New York: College Board.

Yang J., McCrae, R. R., Costa, P. T., Jr., Dai X., Yao S., Cai, T., & Gao, B. (1999). Cross-cultural personality assessment in psychiatric populations: The NEO-PI-R in the People's Republic of China. *Psychological Assessment, 11,* 359–368.

Yates, B. T., & Taub, J. (2003). Assessing the costs, benefits, cost-effectiveness, and cost-benefit of psychological assessment: We should, we can, and here's how. *Psychological Assessment, 15,* 478–495.