

Evidence-Based Assessment

John Hunsley¹ and Eric J. Mash²

¹School of Psychology, University of Ottawa, Ottawa, Ontario, K1N 6N5 Canada; email: hunch@uottawa.ca

²Department of Psychology, University of Calgary, Calgary, Alberta T2N 1N4 Canada; email: mash@ucalgary.ca

Annu. Rev. Clin. Psychol. 2007. 3:29–51

First published online as a Review in
Advance on October 12, 2006

The *Annual Review of Clinical Psychology* is
online at <http://clinpsy.annualreviews.org>

This article's doi:
10.1146/annurev.clinpsy.3.022806.091419

Copyright © 2007 by Annual Reviews.
All rights reserved

1548-5943/07/0427-0029\$20.00

Key Words

psychological assessment, incremental validity, clinical utility

Abstract

Evidence-based assessment (EBA) emphasizes the use of research and theory to inform the selection of assessment targets, the methods and measures used in the assessment, and the assessment process itself. Our review focuses on efforts to develop and promote EBA within clinical psychology. We begin by highlighting some weaknesses in current assessment practices and then present recent efforts to develop EBA guidelines for commonly encountered clinical conditions. Next, we address the need to attend to several critical factors in developing such guidelines, including defining psychometric adequacy, ensuring appropriate attention is paid to the influence of comorbidity and diversity, and disseminating accurate and up-to-date information on EBAs. Examples are provided of how data on incremental validity and clinical utility can inform EBA. Given the central role that assessment should play in evidence-based practice, there is a pressing need for clinically relevant research that can inform EBAs.

Contents

INTRODUCTION.....	30
EXAMPLES OF CURRENT PROBLEMS AND LIMITATIONS IN CLINICAL ASSESSMENT.....	31
Problems with Some Commonly Taught and Used Tests.....	31
Problems in Test Selection and Inadequate Assessment....	32
Problems in Test Interpretation...	32
Limited Evidence for Treatment Utility of Commonly Used Tests.....	32
DEFINING EVIDENCE-BASED ASSESSMENT OF SPECIFIC DISORDERS/CONDITIONS...	33
Disorders Usually First Diagnosed in Youth.....	34
Anxiety Disorders.....	35
Mood Disorders.....	36
Personality Disorders.....	37
Couple Distress.....	37
CHALLENGES IN DEVELOPING AND DISSEMINATING EVIDENCE-BASED ASSESSMENT.....	38
Defining Psychometric Adequacy..	38
Addressing Comorbidity.....	40
Addressing Diversity.....	40
Dissemination.....	41
INCREMENTAL VALIDITY AND EVIDENCE-BASED ASSESSMENT.....	43
Data from Multiple Informants...	43
Data from Multiple Instruments..	44
CLINICAL UTILITY AND EVIDENCE-BASED ASSESSMENT.....	45
CONCLUSIONS.....	46

INTRODUCTION

Over the past decade, attention to the use of evidence-based practices in health care services has grown dramatically. Developed first

in medicine (Sackett et al. 1996), a number of evidence-based initiatives have been undertaken in professional psychology, culminating with the American Psychological Association policy statement on evidence-based practice in psychology (Am. Psychol. Assoc. Presid. Task Force Evid.-Based Pract. 2006). Although the importance of assessment has been alluded to in various practice guidelines and discussions of evidence-based psychological practice, by far the greatest attention has been on intervention. However, without a scientifically sound assessment literature, the prominence accorded evidence-based treatment has been likened to constructing a magnificent house without bothering to build a solid foundation (Achenbach 2005). In their recent review of clinical assessment, Wood et al. (2002) advanced the position that it is necessary for the field to have assessment strategies that are clinically relevant, culturally sensitive, and scientifically sound. With these factors in mind, the focus of our review is on recent efforts to develop and promote evidence-based assessment (EBA) within clinical psychology.

From our perspective, EBA is an approach to clinical evaluation that uses research and theory to guide the selection of constructs to be assessed for a specific assessment purpose, the methods and measures to be used in the assessment, and the manner in which the assessment process unfolds. It involves the recognition that, even with data from psychometrically strong measures, the assessment process is inherently a decision-making task in which the clinician must iteratively formulate and test hypotheses by integrating data that are often incomplete or inconsistent. A truly evidence-based approach to assessment, therefore, would involve an evaluation of the accuracy and usefulness of this complex decision-making task in light of potential errors in data synthesis and interpretation, the costs associated with the assessment process and, ultimately, the impact the assessment had on clinical outcomes for the person(s) being assessed.

In our review of EBA, we begin by briefly illustrating some current weaknesses and lacunae in clinical assessment activities that underscore why a renewed focus on the evidence base for clinical assessment instruments and activities is necessary. We then present recent efforts to operationalize EBA for specific disorders and describe some of the challenges in developing and disseminating EBA. Finally, we illustrate how a consideration of incremental validity and clinical utility can contribute to EBAs.

EXAMPLES OF CURRENT PROBLEMS AND LIMITATIONS IN CLINICAL ASSESSMENT

In this section, we illustrate some of the ways in which current clinical assessment practices may be inconsistent with scientific evidence. Our brief illustrations are not intended as a general indictment of the value of clinical assessments. Rather, by focusing on some frequently used instruments and common assessment activities, our intent is to emphasize that clients involved in assessments may not always be receiving services that are optimally informed by science. As recipients of psychological services, these individuals deserve, of course, nothing less than the best that psychological science has to offer them.

Problems with Some Commonly Taught and Used Tests

Over the past three decades, numerous surveys have been conducted on the instruments most commonly used by clinical psychologists and taught in graduate training programs and internships. With some minor exceptions, the general patterns have remained remarkably consistent and the relative rankings of specific instruments have changed very little over time (e.g., Piotrowski 1999). Unfortunately, many of the tools most frequently taught and used have either limited or mixed supporting empirical evidence. In the past several years, for example, the Rorschach inkblot

test has been the focus of a number of literature reviews (e.g., Hunsley & Bailey 1999, 2001; Meyer & Archer 2001; Stricker & Gold 1999; Wood et al. 2003). There appears to be general agreement that the test (*a*) must be administered, scored, and interpreted in a standardized manner, and (*b*) has appropriate reliability and validity for at least a limited set of purposes. Beyond this minimal level of agreement, however, there is no consensus among advocates and critics on the evidence regarding the clinical value of the test.

The Rorschach is not the only test for which clinical use appears to have outstripped empirical evidence. In this regard, it is illuminating to contrast the apparent popularity of the Thematic Apperception Test (TAT; Murray 1943) and various human figure drawings tasks, both of which usually appear among the ten most commonly recommended and used tests in these surveys, with reviews of these tests' scientific adequacy. There is evidence that some apperceptive measures can be both reliable and valid (e.g., Spangler 1992): This is not the case with the TAT itself. Decades of research have documented the enormous variability in the manner in which the test is administered, scored, and interpreted. As a result, Vane's (1981) conclusion that no cumulative evidence supports the test's reliability and validity still holds today (Rossini & Moretti 1997). In essence, it is a commonly used test that falls well short of professional standards for psychological tests. The same set of problems besets the various types of projective drawing tests. In a recent review, Lally (2001) concluded that the most frequently researched and used approaches to scoring projective drawings fail to meet legal standards for a scientifically valid technique. Scoring systems for projective drawings emphasizing the frequency of occurrence of multiple indicators of psychopathology fared somewhat better, with Lally (2001) suggesting that "[a]lthough their validity is weak, their conclusions are limited in scope, and they appear to offer no additional information over other psychological tests, it can

Evidence-based assessment (EBA):

the use of research and theory to inform the selection of assessment targets, the methods and measures to be used, and the manner in which the assessment process unfolds and is, itself, evaluated

Evidence-based psychological practice:

the use of best available evidence to guide the provision of psychological services, while taking into account both a clinician's expertise and a client's context and values

Assessment purposes:

psychological assessment can be conducted for a number of purposes (e.g., diagnosis, treatment evaluation), and a measure's psychometric properties pertaining to one purpose may not generalize to other purposes

Incremental validity:

the extent to which additional data contribute to the prediction of a variable beyond what is possible with other sources of data

Clinical utility: the extent to which the use of assessment data leads to demonstrable improvements in clinical services and, accordingly, results in improvements in client functioning

TAT: Thematic Apperception Test

IQ: intelligence quotient

at least be argued that they cross the relatively low hurdle of admissibility” (p. 146).

Problems in Test Selection and Inadequate Assessment

The results of clinical assessments can often have a significant impact on those being assessed. Nowhere is this more evident than in evaluations conducted for informing child custody decisions. As a result, numerous guidelines have been developed to assist psychologists in conducting sound child custody evaluations (e.g., Am. Psychol. Assoc. 1994). It appears, however, that psychologists often fail to follow these guidelines or to heed the cautions contained within them. For example, a survey of psychologists who conduct child custody evaluations found that projective tests were often used to assess child adjustment (Ackerman & Ackerman 1997). As we described above, apperceptive tests, projective drawings, and other projectives often do not possess evidence of their reliability and validity. Moving beyond self-report information of assessment practices, Horvath et al. (2002) conducted content analyses of child custody evaluation reports included in court records. They found considerable variability in the extent to which professional guidelines were followed. For example, evaluators often failed to assess general parenting abilities and the ability of each parent to meet his/her child’s needs. The assessment of potential domestic violence and child abuse was also frequently found to be neglected by evaluators.

Problems in Test Interpretation

Because of the care taken in developing norms and establishing reliability and validity indices, the Wechsler intelligence scales are typically seen as among the psychometrically strongest psychological instruments available. Interpretation of the scales typically progresses from a consideration of the full-scale IQ score, to the verbal and performance IQ scores, and then to the factor scores

(e.g., Groth-Marnat 2003). The availability of representative norms and supporting validity studies provide a solid foundation for using these scores to understand a person’s strengths and weaknesses in the realm of mental abilities.

It is also common, however, for authorities to recommend that the next interpretive step involve consideration of the variability between and within subtests (e.g., Flanagan & Kaufman 2004, Kaufman & Lichtenberger 1999). There are, however, a number of problems with this practice. First, the internal consistency of each subtest is usually much lower than that associated with the IQ and factor scores. This low reliability translates into reduced precision of measurement, which leads directly to an increased likelihood of false positive and false negative conclusions about the ability measured by the subtest. Second, there is substantial evidence over several decades that the information contained in subtest profiles adds little to the prediction of either learning behaviors or academic achievement once the IQ scores and factor scores are taken into account (Watkins 2003). An evidence-based approach to the assessment of intelligence would indicate that nothing is to be gained, and much is to be potentially lost, by considering subtest profiles.

Limited Evidence for Treatment Utility of Commonly Used Tests

In test development and validation, the primary foci have been determining the reliability and validity of an instrument. For example, Meyer et al. (2001) provided extensive evidence that many psychological tests have substantial validity when used for clinically relevant purposes. Assessment, however, is more than the use of one or two tests: It involves the integration of a host of data sources, including tests, interviews, and clinical observations. Unfortunately, despite the compelling psychometric evidence for the validity of many psychological tests, almost no research addresses the accuracy (i.e., validity) or the

usefulness (i.e., utility) of psychological assessments (Hunsley 2002, Smith 2002).

In particular, as numerous authors have commented over the years (e.g., Hayes et al. 1987), surprisingly little attention has been paid to the treatment utility of commonly used psychological instruments and methods. Although diagnosis has some utility in determining the best treatment options for clients, there is a paucity of evidence on the degree to which clinical assessment contributes to beneficial treatment outcomes (Nelson-Gray 2003). A recent study by Lima et al. (2005) illustrates the type of utility information that can and should be obtained about assessment tools. These researchers had clients complete the Minnesota Multiphasic Personality Inventory-2 (MMPI-2) prior to commencing treatment; half the treating clinicians received feedback on their client's MMPI-2 data, half did not. Clients presented with a range of diagnoses, with the most common being mood disorders, anxiety disorders, substance-related disorders, adjustment disorders, eating disorders, and personality disorders. Between-group comparisons were conducted on variables related to treatment outcome. In sum, the researchers found that providing clinicians with these results as a potential aid in treatment planning had no positive impact on variables such as improvement ratings or premature termination rates. These data provide evidence that utility, even from an instrument as intensively researched as the MMPI-2, should not be assumed.

DEFINING EVIDENCE-BASED ASSESSMENT OF SPECIFIC DISORDERS/CONDITIONS

In light of the frequent discrepancies between the research base of an assessment instrument and the extent and manner of its use in clinical practice, the need for evidence-based assessment practices is obvious. From our perspective, three critical aspects should define EBA (Hunsley & Mash 2005, Mash & Hunsley 2005). First, research findings and

scientifically viable theories on both psychopathology and normal human development should be used to guide the selection of constructs to be assessed and the assessment process. Second, as much as possible, psychometrically strong measures should be used to assess the constructs targeted in the assessment. Specifically, these measures should have replicated evidence of reliability, validity, and, ideally, clinical utility. Given the range of purposes for which assessment instruments can be used (e.g., screening, diagnosis, treatment monitoring) and the fact that psychometric evidence is always conditional (based on sample characteristics and assessment purpose), supporting psychometric evidence must be available for each purpose for which an instrument or assessment strategy is used. Psychometrically strong measures must also possess appropriate norms for norm-referenced interpretation and/or replicated supporting evidence for the accuracy (i.e., sensitivity, specificity, predictive power, etc.) of cut-scores for criterion-referenced interpretation. Third, although at present little evidence bears on the issue, it is critical that the entire process of assessment (i.e., selection, use, and interpretation of an instrument, and integration of multiple sources of assessment data) be empirically evaluated. In other words, a critical distinction must be made between evidence-based assessment methods and tools, on the one hand, and evidence-based assessment processes, on the other.

In 2005, special sections in two journals, *Journal of Clinical Child and Adolescent Psychology* and *Psychological Assessment*, were devoted to developing EBA guidelines, based on the aforementioned principles, for commonly encountered clinical conditions. As many authors in these special sections noted, despite the voluminous literature on psychological tests relevant to clinical conditions, few concerted attempts have been made to draw on the empirical evidence to develop assessment guidelines, and even fewer evaluations of the utility of assessment guidelines. In the following sections we summarize key points authors

Treatment utility:

the extent to which assessment methods and measures contribute to improvement in the outcomes of psychological treatments

MMPI-2:

Minnesota Multiphasic Personality Inventory-2

ADHD: attention-deficit/hyperactivity disorder

raised about the evidence-based assessment of disorders usually first diagnosed in youth, anxiety disorders, mood disorders, personality disorders, and couple distress.

Disorders Usually First Diagnosed in Youth

Pelham et al. (2005) addressed the assessment of attention-deficit/hyperactivity disorder (ADHD). Based on their literature review, they contended that data obtained from symptom rating scales, completed by both parent and teacher, provide the best information for diagnostic purposes. Despite the widespread use of structured and semistructured interviews, the evidence indicates that they have no incremental validity or utility once data from brief symptom rating scales are considered. Moreover, the authors argued that the diagnostic assessment itself has little treatment utility, especially as the correlation between ADHD symptoms and functional impairment is modest. Accordingly, they suggested that a full assessment of impairments and adaptive skills should be the priority once diagnosis is established. This would involve assessment of (a) functioning in specific domains known to be affected by the disorder (peer relationships, family environment, and school performance) and (b) specific target behaviors that will be directly addressed in any planned treatment.

Drawing on extensive psychopathology research on externalizing behaviors such as oppositional behavior, aggression, physical destructiveness, and stealing, McMahon & Frick (2005) recommended that the evidence-based assessment of conduct problems focus on (a) the types and severity of the conduct problems and (b) the resulting impairment experienced by the youth. Clinicians should also obtain information on the age of onset of severe conduct problems. Compared with an onset after the age of 10 years, onset before the age of 10 is associated with more extreme problems and a greater likelihood of subsequent antisocial and criminal acts. The influence of temperament and social environment may also dif-

fer in child- versus adolescent-onset conduct problems. Many other conditions may co-occur, especially ADHD, depression, and anxiety disorders. Screening for these conditions, therefore, typically is warranted. A range of behavior rating scales, semistructured interviews, and observational systems are available to obtain data on primary and associated features of youth presenting with conduct problems and, as with most youth disorders, obtaining information from multiple informants is critical. However, as McMahon & Frick (2005) indicated, many of these measures are designed for diagnostic and case conceptualization purposes—few have been examined for their suitability in tracking treatment effects or treatment outcome.

Based on recent assessment practice parameters and consensus panel guidelines, Ozonoff et al. (2005) outlined a core battery for assessing autism spectrum disorders. The battery consisted of a number of options for assessing key aspects of the disorders, including diagnostic status, intelligence, language skills, and adaptive behavior. As with the assessment of many disorders, some excellent measurement tools developed in research settings have yet to find their way into clinical practice. Moreover, the authors noted that there has been little attempt to conduct research that directly compares different instruments that assess the same domain, thus leaving clinicians with little guidance about which instrument to use. Ozonoff et al. (2005) also made suggestions for the best evidence-based options for assessing additional domains commonly addressed in autism spectrum disorders evaluations, including attention, executive functions, academic functioning, psychiatric comorbidities, environmental context (i.e., school, family, and community), and response to treatment. Importantly, though, they noted that there was no empirical evidence on whether assessing these domains adds meaningfully to the information available from the recommended core battery.

In their analysis of the learning disabilities assessment literature, Fletcher et al.

(2005) emphasized that approaches to classifying learning disabilities and the measurement of learning disabilities are inseparably connected. Accordingly, rather than focus on specific measures used in learning disability evaluations, these authors highlighted the need to evaluate the psychometric evidence for different models of classification/measurement. Four models were reviewed, including models that emphasized (*a*) low achievement, (*b*) discrepancies between aptitude and achievement, (*c*) intraindividual differences in cognitive functioning, and (*d*) responsiveness to intervention. On the basis of the scientific literature, Fletcher and colleagues concluded that (*a*) the low-achievement model suffers from problems with measurement error and, thus, reliability, (*b*) the discrepancy model has been shown in recent meta-analyses to have very limited validity, (*c*) the intraindividual-differences model also suffers from significant validity problems, and (*d*) the response-to-intervention model has demonstrated both substantial reliability and validity in identifying academic underachievers, but is insufficient for identifying learning disabilities. As a result, they recommended that a hybrid model, combining features of the low-achievement and response-to-treatment models, be used to guide the assessment of learning disabilities. Regardless of the ultimate validity and utility of this hybrid model, Fletcher and colleagues' analysis is extremely valuable for underscoring the need to consider and directly evaluate the manner in which differing assumptions about a disorder may influence measurement.

Anxiety Disorders

Two articles in the special sections dealt with a broad range of anxiety disorders. Silverman & Ollendick (2005) reviewed the literature on the assessment of anxiety and anxiety disorder in youth, and Antony & Rowa (2005) reviewed the comparable literature in adults. Both reviews noted that, regardless of age, the high rates of comorbidity among anxiety disorders

pose a significant challenge for clinicians attempting to achieve an accurate and complete assessment. Additionally, because substantial evidence suggests that individuals diagnosed with an anxiety disorder and another psychiatric disorder (such as ADHD or a mood disorder) are more severely impaired than are individuals presenting with either disorder on its own, assessing for the possible presence of another disorder must be a key aspect of any anxiety disorder evaluation.

Silverman & Ollendick (2005) provided an extensive review of instruments available for youth anxiety disorders, including numerous diagnostic interviews, self-report symptom scales, informant symptom-rating scales (including parent, teacher, and clinician), and observational tasks. Although psychometrically sound instruments are available, many obstacles face clinicians wishing to conduct an evidence-based assessment. For example, the authors reported that efforts to accurately screen for the presence of an anxiety disorder may be hampered by the fact that scales designed to measure similar constructs have substantially different sensitivity and specificity properties. Another obstacle noted by the authors is that all of the measures used to quantify symptoms and anxious behaviors rely on an arbitrary metric (cf. Blanton & Jaccard 2006, Kazdin 2006). As a result, we simply do not know how well scores on these measures map on to actual disturbances and functional impairments. A final example stems from the ubiquitous research finding that youth and their parents are highly discordant in their reports of anxiety symptoms. In light of such data, it is commonly recommended that both youth and parent reports be obtained, but, as Silverman & Ollendick (2005) cautioned, care must be exercised to ensure that neither is treated as a gold standard when diagnosing an anxiety disorder.

Antony & Rowa (2005) emphasized the importance of assessing key dimensions that cut across anxiety disorders in their review of the adult anxiety disorder literature. Based on diagnostic criteria, anxiety disorders research,

Comorbidity: the co-occurrence of multiple disorders or clinically significant patterns of dysfunction

and expert consensus statements, they recommended that evidence-based assessment for anxiety disorders should target anxiety cues and triggers, avoidance behaviors, compulsions and overprotective behaviors, physical symptoms and responses, comorbidity, skills deficits, functional impairment, social environment factors, associated health issues, and disorder development and treatment history. To illustrate how these dimensions could be assessed, the authors presented an assessment protocol to be used for assessing treatment outcome in the case of panic disorder with agoraphobia. The literature on the assessment of anxiety problems in adults has an abundance of psychometrically strong interviews and self-report measures, and numerous studies have supported the value of obtaining self-monitoring data and using behavioral tests to provide observable evidence of anxiety and avoidance. Nevertheless, Antony & Rowa (2005) cautioned that, even for well-established measures, little validity data exist beyond evidence of how well one measure correlates with another. Echoing the theme raised by Silverman & Ollendick (2005), they emphasized that little is currently known about how well an instrument correlates with responses in anxiety-provoking situations.

Mood Disorders

Three articles in the special sections dealt with mood disorders. Klein et al. (2005) addressed the assessment of depression in youth, Joiner et al. (2005) dealt with the assessment of depression in adults, and Youngstrom et al. (2005) discussed initial steps toward an evidence-based assessment of pediatric bipolar disorder. With respect to the assessment of depression, both sets of authors recommended that the best assessment practice would be to use a validated semistructured interview to address diagnostic criteria, comorbidity, disorder course, family history, and social environment. Additionally, these authors stressed the critical need to include a sensitive and thorough assessment of suicidal ideation

and the potential for suicidal behavior in all depression-related assessments.

Klein et al. (2005) reported that psychometrically strong semistructured interviews are available for the assessment of depression in youth. The need for input from multiple informants is especially important, as younger children may not be able to comment accurately on the time scale associated with depressive experiences or on the occurrence and duration of previous episodes. As with anxiety disorders, there is consistent evidence regarding the limited agreement between youth and parent reports of depression, although recent evidence shows that youth, parent, and teacher reports can all contribute uniquely to the prediction of subsequent outcomes. On the other hand, Klein and colleagues (2005) cautioned that depressed parents have been found to have a lower threshold, relative to nondepressed parents, in identifying depression in their children. Ratings scales, for both parents and youth, were described as especially valuable for the assessment purposes of screening, treatment monitoring, and treatment evaluation. Unfortunately, most such rating scales have rather poor discriminant validity, especially with respect to anxiety disorders. The authors also indicated that, because so little research exists on the assessment of depression in preschool-age children, it is not possible to make strong recommendations for clinicians conducting such assessments.

Based on extensive research, Joiner et al. (2005) concluded that depression can be reliably and validly assessed, although they cautioned that attention must be paid to the differential diagnosis of subtypes of depression, such as melancholic-endogenous depression, atypical depression, and seasonal affective disorder. The authors noted that there is no strong evidence of gender or ethnicity biases in depression assessment instruments and, although there is some concern about inflated scores on self-report measures among older adults (primarily due to items dealing with somatic and vegetative symptoms), good measures are available for use with older

adults. Psychometrically strong measures exist for both screening and treatment monitoring purposes; for this latter purpose, some research has indicated that clinician ratings are more sensitive to treatment changes than are client ratings. Nevertheless, Joiner and colleagues emphasized that the methods and measures currently available to assess depression have yet to demonstrate their value in the design or delivery of intervention services to depressed adults.

Because of the relative recency of the research literature on bipolar disorder in youth and the ongoing debate about the validity of the diagnosis in children and adolescents, Youngstrom et al. (2005) focused on providing guidance on how evidence-based assessment might develop for this disorder. The paucity of psychometrically adequate interviews and self-report instruments led the authors to address foundational elements that should be included in an assessment. For example, they stressed the need to carefully consider family history: Although an average odds ratio of 5 has been found for the risk of the disorder in a child if a parent has the disorder, approximately 95% of youth with a parent who has bipolar disorder will not, themselves, meet diagnostic criteria. They also emphasized the importance of attending to symptoms that are relatively specific to the disorder (e.g., elevated mood, grandiosity, pressured speech, racing thoughts, and hypersexuality) and to evidence of patterns such as mood cycling and distinct spontaneous changes in mood states. Because of the likely lack of insight or awareness in youth of possible manic symptoms, collateral information from teachers and parents has been shown to be particularly valuable in predicting diagnostic status. Finally, due to the need to identify patterns of mood shifts and concerns about the validity of retrospective recall, Youngstrom and colleagues (2005) strongly recommended that an assessment for possible bipolar disorder should occur over an extended period, thus allowing the clinician an opportunity to obtain data from repeated evaluations.

Personality Disorders

In their review of the literature on the assessment of personality disorders, Widiger & Samuel (2005) described several scientifically sound options for both semistructured interviews and self-report measures, although they did note that not all instruments have normative data available. To maximize accuracy and minimize the burden on the clinician, they recommended a strategy whereby a positive response on a self-report instrument is followed up with a semistructured interview. Concerns about limited self-awareness and inaccuracies in self-perception among individuals being assessed for a personality disorder raise the issue of relying on self-report, whether on rating scales or interviews. Moreover, given the potential for both gender and ethnicity biases to occur in these instruments, clinicians must be alert to the possibility of diagnostic misclassification. As with youth disorders, the use of collateral data is strongly encouraged, especially as research has indicated that both client and informant provide data that contribute uniquely to a diagnostic assessment. Widiger & Samuel (2005) also underscored the need for the development of measures to track treatment-related changes in maladaptive personality functioning.

Couple Distress

Snyder et al. (2005) presented a conceptual framework for assessing couple functioning that addresses both individual and dyadic characteristics. Drawing on extensive studies of intimate relationships, they highlighted the need to assess relationship behaviors (e.g., communication, handling conflict), relationship cognitions (e.g., relationship-related standards, expectations, and attributions), relationship affect (e.g., rates, duration, and reciprocity of both negative and positive affect), and individual distress. Much valuable information on these domains can be obtained from psychometrically sound rating scales; however, the authors concluded that

most self-report measures have not undergone sufficient psychometric analysis to warrant their clinical use. Moreover, the authors noted that very little progress had been made in developing interview protocols that demonstrate basic levels of reliability and validity. They also stressed the unique contribution afforded by the use of analog behavior observation in assessing broad classes of behavior such as communication, power, problem solving, and support/intimacy. Thus, rather than recommending a specific set of measures to be used in assessing couples, Snyder and colleagues (2005) suggested that their behavior/cognition/affect/distress framework be used to guide the selection of constructs and measures as the assessment process progressed from identifying broad relationship concerns to specifying elements of these concerns that are functionally linked to the problems in couple functioning.

CHALLENGES IN DEVELOPING AND DISSEMINATING EVIDENCE-BASED ASSESSMENT

Based on the foregoing analysis of EBA for commonly encountered clinical conditions, many scientific and logistic challenges must be addressed. In this section, we focus on some of the more pressing issues stemming from efforts to develop a truly evidence-based approach to assessment in clinical psychology. We begin with the basic question of what constitutes “good enough” psychometric criteria, then move on to examine issues such as comorbidity, attention to diversity parameters, and the promotion of EBA in clinical practice. Additional potential challenges in EBA, such as the use of multiple measures and multiple informants and the integration of assessment data, are discussed below in a section on incremental validity.

Defining Psychometric Adequacy

In their presentation on the assessment of depression in youth, Klein and colleagues (2005)

queried what constitutes an acceptable level of reliability or validity in an instrument. After many decades of research on test construction and evaluation, it would be tempting to assume that the criteria for what constitutes “good enough” evidence to support the clinical use of an instrument have been clearly established: Nothing could be further from the case. The *Standards for Educational and Psychological Testing* (Am. Educ. Res. Assoc., Am. Psychol. Assoc., Natl. Council. Meas. Educ. 1999) set out generic standards to be followed in developing and using tests, and these standards are well accepted by psychologists. In essence, for an instrument to be psychometrically sound, it must be standardized, have relevant norms, and have appropriate levels of reliability and validity (cf. Hunsley et al. 2003). The difficulty comes in defining what standards should be met when considering these characteristics.

As we and many others have stressed in our work on psychological assessment, psychometric characteristics are not properties of an instrument per se, but rather are properties of an instrument when used for a specific purpose with a specific sample. For this reason, many assessment scholars and psychometricians are understandably reluctant to provide precise standards for the psychometric properties that an instrument or strategy must have in order to be used for assessment purposes (e.g., Streiner & Norman 2003). On the other hand, both researchers and clinicians are constantly faced with the decision of whether an instrument is good enough for the assessment task at hand.

Some attempts have been made over the past two decades to delineate criteria for measure selection and use. Robinson, Shaver & Wrightsman (1991) developed evaluative criteria for the adequacy of attitude and personality measures, covering the domains of theoretical development, item development, norms, interitem correlations, internal consistency, test-retest reliability, factor analytic results, known groups validity, convergent validity, discriminant validity, and freedom

from response sets. Robinson and colleagues (1991) also used specific criteria for many of these domains. For example, a coefficient α of 0.80 was deemed exemplary, as was the availability of three or more studies showing the instrument had results that were independent of response biases. More recently, efforts have been made to establish general psychometric criteria for determining disability in speech-language disorders (Agency Healthc. Res. Qual. 2002) and reliability criteria for a multinational measure of psychiatric services (Schene et al. 2000). Taking a different approach, expert panel ratings were used by the Measurement and Treatment Research to Improve Cognition in Schizophrenia Group to develop a consensus battery of cognitive tests to be used in clinical trials in schizophrenia (MATRICS 2006). Rather than specify precise psychometric criteria, panelists were asked to rate, on a nine-point scale, each proposed test's characteristics, including test-retest reliability, utility as a repeated measure, relation to functional outcome, responsiveness to treatment change, and practicality/tolerability.

In a recent effort to promote the development of EBA in clinical assessment, we developed a rating system for instruments that was intended to embody a "good enough" principle across psychometric categories with clear clinical relevance (Hunsley & Mash 2006). We focused on nine categories: norms, internal consistency, interrater reliability, test-retest reliability, content validity, construct validity, validity generalization, sensitivity to treatment change, and clinical utility. Each of these categories is applied in relation to a specific assessment purpose (e.g., case conceptualization) in the context of a specific disorder or clinical condition (e.g., eating disorders, self-injurious behavior, and relationship conflict). For each category, a rating of acceptable, good, excellent, or not applicable is possible. The precise nature of what constitutes acceptable, good, and excellent varied, of course, from category to category. In general, though, a rating of acceptable indi-

cated that the instrument meets a minimal level of scientific rigor, good indicated that the instrument would generally be seen as possessing solid scientific support, and excellent indicated there was extensive, high-quality supporting evidence. When considering the clinical use of a measure, it would be desirable to use only those measures that would meet, at a minimum, our criteria for good. However, as measure development is an ongoing process, we felt it was important to provide the option of the acceptable rating in order to fairly evaluate (a) relatively newly developed measures and (b) measures for which comparable levels of research evidence are not available across all psychometric categories in our rating system.

To illustrate this rating system, we focus on the internal consistency category. Although a number of indices of internal consistency are available, α is the most widely used index (Streiner 2003). Therefore, even though concerns have been raised about the potential for undercorrection of measurement error with this index (Schmidt et al. 2003), we established criteria for α in our system. Across all three possible ratings in the system, we encouraged attention to the preponderance of research results. Such an approach allows a balance to be maintained between (a) the importance of having replicated results and (b) the recognition that variability in samples and sampling strategies will yield a range of reliability values for any measure. Ideally, meta-analytic indices of effect size could be used to provide precise estimates from the research literature. Recommendations for what constitutes good internal consistency vary from author to author, but most authorities seem to view 0.70 as the minimum acceptable value (cf. Charter 2003). Accordingly, our rating of adequate is appropriate when the preponderance of evidence indicated values of 0.70–0.79. For a rating of good, we required that the preponderance of evidence indicated values of 0.80–0.89. Finally, because of cogent arguments that an α value of at least 0.90 is highly desirable in clinical assessment contexts (Nunnally &

Bernstein 1994), we required that the preponderance of evidence indicated values ≥ 0.90 for an instrument to be rated as having excellent internal consistency. That being said, it is also possible for α to be too (artificially) high, as a value close to unity typically indicates substantial redundancy among items.

Addressing Comorbidity

As stressed by all contributors to the special sections on EBA described above, the need to assess accurately comorbid conditions is a constant clinical reality. Simply put, people seen in clinical settings, across the age span, frequently meet diagnostic criteria for more than one disorder or have symptoms from multiple disorders even if they occur at a subclinical level (Kazdin 2005). Indeed, recent nationally representative data on comorbidity in adults indicated that 45% of those meeting criteria for an anxiety, mood, impulse control, or substance disorder also met criteria for one or two additional diagnoses (Kessler et al. 2005). At present, it is not possible to disentangle the various factors that may account for this state of affairs. True heterogeneity among the patterns of presenting symptoms, poor content validity within some symptom measures, limitations inherent in current diagnostic categories, and the use of mixed-age samples to estimate lifetime prevalence of comorbidity, singly or together, can contribute to the high observed rates of comorbidity (Achenbach 2005, Kraemer et al. 2006).

However, evidence is emerging that observed comorbidity is at least partially due to the presence of core pathological processes that underlie the overt expression of a seemingly diverse range of symptoms (Krueger & Markon 2006, Widiger & Clark 2000). In particular, the internalizing and externalizing dimensions first identified as relevant to childhood disorders appear to have considerable applicability to adult disorders. In a cross-cultural study examining the structure of psychiatric comorbidity in 14 countries, Krueger et al. (2003) found that a two-factor (inter-

nalizing and externalizing) model accurately represented the range of commonly reported symptoms. Moreover, there is evidence that individuals diagnosed with comorbid conditions are more severely impaired in daily life functioning, are more likely to have a chronic history of mental health problems, have more physical health problems, and use more health care services than do those with a single diagnosable condition (Newman et al. 1998). Hence, for numerous reasons, the evidence-based assessment of any specific disorder requires that the presence of commonly encountered comorbid conditions, as defined by the results of extant psychopathology research, be evaluated.

Fortunately, viable options exist for conducting such an evaluation. Conceptualizing the assessment process as having multiple, interdependent stages, it is relatively straightforward to have the initial stage address more general considerations such as a preliminary broadband evaluation of symptoms and life context. As indicated by many contributors to the special sections, some semistructured interviews provide such information, for both youth and adults. However, time constraints and a lack of formal training, among other considerations, may leave many clinicians disinclined to use these instruments. Good alternatives do exist, including multidimensional screening tools and brief symptom checklists for disorders most frequently comorbid with the target disorder (Achenbach 2005, Mash & Hunsley 2005). Additionally, it may be worthwhile to ensure that the assessment includes an evaluation of common parameters or domains that cut across the comorbid conditions. For example, regardless of the specific diagnoses being evaluated, situational triggers and avoidance behaviors are particularly important in the EBA of anxiety disorders in adults (Antony & Rowa 2005).

Addressing Diversity

When considering the applicability and potential utility of assessment instruments for a

particular clinical purpose with a specific individual, clinicians must attend to diversity parameters such as age, gender, and ethnicity. Dealing with developmental differences, throughout the life span, requires measures that are sensitive to developmental factors and age-relevant norms. Unfortunately, it is often the case that measures for children and adolescents are little more than downward extensions of those developed for use with adults (Silverman & Ollendick 2005). On the other hand, relevant research often is available to guide the clinician's choice of variables to assess. For example, research indicating that girls are more likely to use indirect and relational forms of aggression than they are physical aggression may point to different assessment targets for girls and boys when assessing youth suspected of having a conduct disorder (Crick & Nelson 2002). Likewise, the literature is rapidly expanding on ethnic and cultural variability in symptom expression and the psychometric adequacy of commonly used self-report measures in light of this variability (e.g., Achenbach et al. 2005, Jones et al. 2000).

Presentations of the conceptual and methodological requirements and challenges involved in translating a measure into a different language and determining the appropriateness of a measure and its norms to a specific cultural group are also widely available (e.g., Geisinger 1994, van Widenfelt et al. 2005). As described succinctly by Snyder et al. (2005), four main areas need to be empirically evaluated in using or adapting instruments cross-culturally. These are (a) linguistic equivalence of the measure, (b) psychological equivalence of items, (c) functional equivalence of the measure, including predictive and criterion validity, and (d) scalar equivalence, including regression line slope and comparable metrics. Addressing these areas provides some assurance that cultural biases have been minimized or eliminated from a measure. However, many subtle influences may impede efforts to develop culturally appropriate measures. For example, investigations into the associations be-

tween ethnicity/culture and scores on a measure must be sensitive to factors such as subgroup differences in cultural expression and identity, socioeconomic status, immigration and refugee experiences, and acculturation (Alvidrez et al. 1996).

Notwithstanding the progress made in developing assessment methods and measures that are sensitive to diversity considerations, a very considerable challenge remains in developing EBAs that are sensitive to aspects of diversity. As cogently argued by Kazdin (2005), the number of potential moderating variables is so large that it is simply not realistic to expect that we will be able to develop an assessment evidence base that fully encompasses the direct and interactional influences these variables have on psychological functioning. Therefore, in addition to attending to diversity parameters in designing, conducting, and interpreting assessment research, psychologists need to be able to balance knowledge of instrument norms with an awareness of an individual's characteristics and circumstances. The availability, for commonly used standardized instruments, of nationally representative norms that are keyed to gender and age has great potential for aiding clinicians in understanding client functioning (Achenbach 2005). Such data must, however, be augmented with empirically derived principles that can serve as a guide in determining which elements of diversity are likely to be of particular importance for a given clinical case or situation.

Dissemination

For those interested in advancing the use of EBAs, the situation is definitely one in which the "glass" can either be seen as "half full" or as "half empty." Recent surveys of clinical psychologists indicate that a relatively limited amount of professional time is devoted to psychological assessment activities (Camara et al. 2000) and that relatively few clinical psychologists routinely formally evaluate treatment outcome (Cashel 2002, Hatfield &

Ogles 2004). Despite mounting pressure for the use of outcome assessment data in developing performance indicators and improving service delivery (e.g., Heinemann 2005), one recent survey found that, even when outcome assessments were mandatory in a clinical setting, most clinicians eschewed the available data and based their practices on an intuitive sense of what they felt clients needed (Garland et al. 2003). Furthermore, the assessment methods and measures typically taught in graduate training programs and those most frequently used by clinical psychologists bear little resemblance to the methods and measures involved in EBAs (Hunsley et al. 2004). On the other hand, an increasing number of assessment volumes wholeheartedly embrace an evidence-based approach (e.g., Antony & Barlow 2002, Hunsley & Mash 2006, Mash & Barkley 2006, Nezu et al. 2000).

In 2000, the American Psychiatric Association published the *Handbook of Psychiatric Measures*, a resource designed to offer information to mental health professionals on the availability of self-report measures, clinician checklists, and structured interviews that may be of value in the provision of clinical services. This compendium includes reviews of both general (e.g., health status, quality of life, family functioning) and diagnosis-specific assessment instruments. For each measure, there is a brief summary of its intended purpose, psychometric properties, likely clinical utility, and the practical issues encountered in using the measure. Relatedly, in their continuing efforts to improve the quality of mental health assessments, the American Psychiatric Association just released the second edition of its *Practice Guideline for the Psychiatric Evaluation of Adults* (2006). Drawing from both current scientific knowledge and the realities of clinical practice, this guideline addresses both the content and process of an evaluation. Despite the longstanding connection between psychological measurement and the profession of clinical psychology, no comparable concerted effort has been made to develop assessment guidelines in professional psychology. From a

purely professional perspective, this absence of leadership in the assessment field seems unwise, but only time will tell whether the lack of involvement of organized psychology proves detrimental to the practice of psychological assessment.

Indications are growing that, across professions, clinicians are seeking assessment tools they can use to determine a client's level of pretreatment functioning and to develop, monitor, and evaluate the services received by the client (Barkham et al. 2001, Bickman et al. 2000, Hatfield & Ogles 2004); in other words, exactly the type of data encompassed by EBAs. Assuming, therefore, that at least a sizeable number of clinical psychologists might be interested in adopting EBA practices, what might be some of the issues that must be confronted if widespread dissemination is to occur? There must be some consensus on the psychometric qualities necessary for an instrument to merit its use in clinical services and, ideally, there should be consensus among experts about instruments that possess those qualities when used for a specific assessment purpose (Antony & Rowa 2005). Although some clinicians may simply be unwilling to adopt new assessment practices, it is imperative that the response cost for those willing to learn and use EBAs be relatively minimal. It would be ideal if most measures used in EBAs were brief, inexpensive to use, had robust reliability and validity characteristics across client groups, and were straightforward to administer, score, and interpret. To enhance the value of any guideline developed for EBAs, guidelines would need to be succinct, employ presentational strategies to depict complex psychometric data in a straightforward manner, be easily accessible (e.g., downloadable documents on a Web site), and be regularly updated as warranted by advances in research (Mash & Hunsley 2005). The strategies, methods, and technologies needed to develop and maintain such guidelines are all readily available. For example, meta-analytic summaries using data across studies can provide accurate estimates of the psychometric characteristics of

a measure. The challenge is to pull together all the requisite components in a scientifically rigorous and sustainable fashion.

One final point is also clear: Simply doing more of the same in terms of the kind of assessment research typically conducted will not advance the use of EBAs. At present, there continues to be a proliferation of measures, the usual study focuses on a measure's concurrent validity with respect to other similar measures, and relatively little attention is paid to the prediction of clients' real-world functioning or to the clinical usefulness of the measure (Antony & Rowa 2005, Kazdin 2005, McGrath 2001). All of these mitigate against the likelihood of clinicians having access to scientifically solid assessment tools that are both clinically feasible and useful. In the sections below, we turn to the clinical features necessary to the uptake of EBAs—namely, incremental validity and utility.

INCREMENTAL VALIDITY AND EVIDENCE-BASED ASSESSMENT

Incremental validity is essentially a straightforward concept that addresses the question of whether data from an assessment tool add to the prediction of a criterion beyond what can be accomplished with other sources of data (Hunsley & Meyer 2003, Sechrest 1963). Nested within this concept, however, are a number of interrelated clinical questions that are crucial to the development and dissemination of EBA. These include questions such as whether it is worthwhile, in terms of both time and money, to (a) use a given instrument, (b) obtain data on the same variable using multiple methods, (c) collect parallel information from multiple informants, and (d) even bother collecting assessment data beyond information on diagnostic status, as most evidence-based treatments are keyed to diagnosis.

Ideally, incremental validity research should be able to provide guidance to clinicians on what could constitute the necessary scope for a given assessment purpose. Such

guidance is especially important in the realm of clinical services to children and adolescents, as the norm for many years has been to collect data on multiple measures from multiple informants. In reality, however, there is little replicated evidence in the clinical literature on which to base such guidance (cf. Garb 2003, Johnston & Murray 2003). This is primarily due to the extremely limited use of research designs and analyses relevant to the question of incremental validity of instruments or data sources. Haynes & Lench (2003) reported, for example, that over a five-year period, only 10% of manuscripts submitted for possible publication in the journal *Psychological Assessment* considered a measure's incremental validity.

Nevertheless, in the literature some important incremental validity data are available that have direct relevance to the practice of clinical assessment. Moreover, a renewed focus on the topic (e.g., Haynes & Lench 2003, McFall 2005) and the availability of guidelines for interpreting what constitutes clinically meaningful validity increments (Hunsley & Meyer 2003) may lead to greater attention to conducting incremental validity analyses. In the following sections, we provide some examples of the ways in which incremental validity research has begun to address commonly encountered challenges in clinical assessment. We begin with research on using data from multiple informants and then turn to the use of data from multiple instruments.

Data from Multiple Informants

As we indicated, in assessing youth, it has been a longstanding practice for clinicians to obtain assessment data from multiple informants such as parents, teachers, and the youths themselves. It is now commonly accepted that, because of differing perspectives, these informant ratings will not be interchangeable but can each provide potentially valuable assessment data (e.g., De Los Reyes & Kazdin 2005). Of course, in usual clinical

services, only a very limited amount of time is available to obtain initial assessment data. The obvious issue, therefore, is determining which informants are optimally placed to provide data most relevant to the assessment question at hand. Several distinct approaches to integrating data from multiple sources can be found in the literature. As summarized by Klein et al. (2005), these include the “or” rule (assuming that the feature targeted in the assessment is present if any informant reports it), the “and” rule (requiring two or more informants to confirm the presence of the feature), and the use of statistical regression models to integrate data from the multiple sources. In the following examples, we focus on the issue of obtaining multiple informant data for the purposes of clinical diagnosis.

Several research groups have been attempting to determine what could constitute a minimal set of assessment activities required for the assessment of ADHD in youth (e.g., Power et al. 1998, Wolraich et al. 2003). Focusing on instruments that are both empirically supported and clinically relevant, Pelham and colleagues (2005) recently synthesized the results of this line of research and drew several conclusions of direct clinical value. First, they concluded that diagnosing ADHD is most efficiently accomplished by relying on data from parent and teacher ADHD rating scales. Structured diagnostic interviews do not possess incremental validity over rating scales and, therefore, are likely to have little value in clinical settings. Second, clinical interviews and/or other rating scales are important for gaining information about the onset of the disorder and ruling out other conditions. Such information can be invaluable in designing intervention services and, consistent with diagnostic criteria, confirmatory data from both teachers and parents are necessary for the diagnosis of ADHD. However, in ruling out an ADHD diagnosis, it is not necessary to use both parent and teacher data: If either informant does not endorse rating

items that possess strong negative predictive power, the youth is unlikely to have a diagnosis of ADHD.

Youngstrom et al. (2004) compared the diagnostic accuracy of six different instruments designed to screen for youth bipolar disorder. Three of these instruments involved parent reports, two involved youth self-reports, and one relied on teacher reports. Parent-based measures consistently outperformed measures based on youth self-report and teacher report in identifying bipolar disorder among youth (as determined by a structured diagnostic interview of the youth and parent). Additionally, the researchers found that no meaningful increment in prediction occurred when data from all measures were combined. Although none of the measures studied was designed to be a diagnostic instrument, and none was sufficient on its own for diagnosing the condition, the clinical implications from these findings are self-evident.

Data from Multiple Instruments

Within the constraints of typical clinical practice, there are perennial concerns regarding which instruments are critical for a given purpose and whether including multiple instruments will improve the accuracy of the assessment. In the research on the clinical assessment of adults, a growing number of studies address these concerns. We have chosen to illustrate what can be learned from this literature by focusing on two high-stakes assessment issues: detecting malingering and predicting recidivism.

The detection of malingering has become an important task in many clinical and forensic settings. Researchers have examined the ability of both specially developed malingering measures and the validity scales included in broadband assessment measures to identify accurately individuals who appear to be feigning clinically significant distress. In this context, Bagby et al. (2005) recently evaluated the

incremental validity of several MMPI-2 validity scales to detect those faking depressive symptoms. Data from the Malingering Depression scale were compared to the F scales in a sample of MMPI-2 protocols that included depressed patients and mental health professionals instructed to feign depression. All validity scales had comparable results in detecting feigned depression, with no one scale being substantially better than any other scale. The Malingering Depression scale was found to have slight, but statistically significant, incremental validity over the other scales. However, the researchers reported that this statistical advantage did not translate into a meaningful clinical advantage: Very few additional “malingering” protocols were accurately identified by the scale beyond what was achieved with the generic validity scales.

The development of actuarial risk scales has been responsible for significant advances in the accurate prediction of criminal recidivism. Seto (2005) examined the incremental validity of four well-established actuarial risk scales, all with substantial empirical support, in predicting the occurrence among adult sex offenders of both serious violent offenses and sexual offense involving physical contact with the victims.

As some variability existed among the scales in terms of their ability to predict accurately both types of criminal offenses, Seto (2005) examined the predictive value of numerous strategies for combining data from multiple scales. These included both the “or” and “and” rules described above, along with strategies that used the average results across scales and statistical optimization methods derived via logistic regression and principal component analysis. Overall, Seto found that no combination of scales improved upon the predictive accuracy of the single best actuarial scale for the two types of criminal offenses. Accordingly, he suggested that evaluators should simply select the single best scale for the assessment purpose, rather than obtaining the information nec-

essary for scoring and interpreting multiple scales.

CLINICAL UTILITY AND EVIDENCE-BASED ASSESSMENT

The concept of clinical utility has received a great deal of attention in recent years. Although definitions vary, an emphasis on garnering evidence regarding actual improvements in both decisions made by clinicians and service outcomes experienced by patients and clients is at the heart of clinical utility, whether the focus is on diagnostic systems (First et al. 2004, Kendell & Jablensky 2003), assessment tools (Hunsley & Bailey 1999, McFall 2005), or intervention strategies (Am. Psychol. Assoc. Presid. Task Force Evid.-Based Pract. 2006).

Without a doubt, recent decades have witnessed considerable advances in the quality and quantity of assessment tools available for studying both normal and abnormal human functioning. On the other hand, despite thousands of studies on the reliability and validity of psychological instruments, very little evidence exists that psychological assessment data have any functional relation to the enhanced provision and outcome of clinical services. Indeed, the Lima et al. (2005) study described above is one of the few examples in which an assessment tool has been examined for evidence of utility. However, a truly evidence-based approach to clinical assessment requires not only psychometric evidence of the soundness of instruments and strategies, but also data on the fundamental question of whether or not the assessment enterprise itself makes a difference with respect to the accuracy, outcome, or efficiency of clinical activities.

One exception to this general state of affairs can be found in the literature on the Outcome Questionnaire-45 (OQ-45). The OQ-45 measures symptoms of distress, interpersonal relations, and social functioning, and has been shown repeatedly to have

OQ-45: Outcome
Questionnaire-45

good psychometric characteristics, including sensitivity to treatment-related change (Vermeersch et al. 2000, 2004). Lambert et al. (2003) conducted a meta-analysis of three large-scale studies (totaling more than 2500 adult clients) in which feedback on session-by-session OQ-45 data from clients receiving treatment was obtained and then, depending on the experimental condition, was either provided or not provided to the treating clinicians. In the clinician feedback condition, the extent of the feedback was very limited, involving simply an indication of whether clients, based on normative data, were making adequate treatment gains, making less-than-adequate treatment gains, or experiencing so few benefits from treatment that they were at risk for negative treatment outcomes. By the end of treatment, in the no-feedback/treatment-as-usual condition, based on OQ-45 data, 21% of clients had experienced a deterioration of functioning and 21% had improved in their functioning. In contrast, in the feedback condition, 13% had deteriorated and 35% had improved. In other words, compared with usual treatment, simply receiving general feedback on client functioning each session resulted in 38% fewer clients deteriorating in treatment and 67% more clients improving in treatment. Such data provide promising evidence of the clinical utility of the OQ-45 for treatment monitoring purposes and, more broadly, of the value of conducting utility-relevant research.

CONCLUSIONS

In this era of evidence-based practice, there is a need to re-emphasize the vital importance of using science to guide the selection and use of assessment methods and instruments. Assessment is often viewed as a clinical activity and service in its own right, but it is important not to overlook the interplay between assessment and intervention that is at the heart of providing evidence-based psychological treatments. This assessment-intervention dialectic, involving the use of assessment data both to plan treatment and to modify the treatment in response to changes in a client's functioning and goals (Weisz et al. 2004), means that EBA has relevance for a broad range of clinical services. Regardless of the purpose of the assessment, the central focus within EBA on the clinical application of assessment strategies makes the need for research on incremental validity and clinical utility abundantly clear. Moreover, although it is fraught with potential problems, the process of establishing criteria and standards for EBA has many benefits, including providing (a) useful information to clinicians on assessment options, (b) indications of where gaps in supporting scientific evidence may exist for currently available instruments, and (c) concrete guidance on essential psychometric criteria for the development and clinical use of assessment instruments.

SUMMARY POINTS

1. Evidence-based assessment (EBA) is a critical, but underappreciated, component of evidence-based practice in psychology.
2. Based on existing research, initial guidelines for EBAs have been delineated for many commonly encountered clinical conditions.
3. Researchers and clinicians are frequently faced with the decision of whether an instrument is good enough for the assessment task at hand. Thus, despite the challenges involved, steps must be taken to determine the psychometric properties that make a measure "good enough" for clinical use.

4. More research is needed on the influence of diversity parameters, such as age, gender, and ethnicity, on assessment methods and measures. Additionally, empirically derived principles that can serve as a guide in determining which elements of diversity are likely to be of particular importance for a given clinical case or situation.
5. A growing body of research addresses the optimal manner for combining data from multiple instruments and from multiple sources. Such research can inform the optimal use of assessment data for various clinical purposes.
6. Clinical utility is emerging as a key consideration in the development of diagnostic systems, assessment, and intervention. The utility of psychological assessment, and of EBA itself, requires much greater empirical attention than has been the case to date.
7. Although a challenging enterprise due to the scope of the assessment literature, the strategies, methods, and technologies needed to develop and maintain EBA guidelines are available and can be used to advance the evidence-based practice of psychology.

LITERATURE CITED

- Achenbach TM. 2005. Advancing assessment of children and adolescents: commentary on evidence-based assessment of child and adolescent disorders. *J. Clin. Child Adolesc. Psychol.* 34:541–47
- Achenbach TM, Rescorla LA, Ivanova MY. 2005. International cross-cultural consistencies and variations in child and adolescent psychopathology. In *Comprehensive Handbook of Multicultural School Psychology*, ed. CL Frisby, CR Reynolds, Cecil R, pp. 674–709. Hoboken, NJ: Wiley
- Ackerman MJ, Ackerman MC. 1997. Custody evaluation practices: a survey of experienced professionals (revisited). *Prof. Psychol. Res. Pract.* 28:137–45
- Agency Healthc. Res. Qual. 2002. Criteria for determining disability in speech-language disorders. *AHRQ Publ. No. 02-E009*. Rockville, MD: AHRQ
- Alvidrez J, Azocar F, Miranda J. 1996. Demystifying the concept of ethnicity for psychotherapy researchers. *J. Consult. Clin. Psychol.* 64:903–8
- Am. Educ. Res. Assoc., Am. Psychol. Assoc., Natl. Counc. Meas. Educ. 1999. *Standards for Educational and Psychological Testing*. Washington, DC: Am. Educ. Res. Assoc. 194 pp.
- Am. Psychiatr. Assoc. 2000. *Handbook of Psychiatric Measures*. Washington, DC: Am. Psychiatr. Publ. 848 pp.
- Am. Psychiatr. Assoc. 2006. *Practice Guideline for the Psychiatric Evaluation of Adults*. 2nd ed. http://www.psych.org/psych_pract/treatg/pg/PsychEval2ePG.04-28-06.pdf
- Am. Psychol. Assoc. 1994. Guidelines for child custody evaluations in divorce proceedings. *Am. Psychol.* 49:677–80
- Am. Psychol. Assoc. Presid. Task Force Evid.-Based Pract. 2006. Evidence-based practice in psychology. *Am. Psychol.* 61:271–85**
- Antony MM, Barlow DH, eds. 2002. *Handbook of Assessment and Treatment Planning for Psychological Disorders*. New York: Guilford
- Antony MM, Rowa K. 2005. Evidence-based assessment of anxiety disorders in adults. *Psychol. Assess.* 17:256–66
- Bagby RM, Marshall MD, Bacchiochi JR. 2005. The validity and clinical utility of the MMPI-2 Malingering Depression scale. *J. Personal. Assess.* 85:304–11

States how evidence-based practices can be operationalized within professional psychology.

Summarizes research on differences in data provided by multiple informants and provides guidance on how to conceptualize and use these different perspectives.

Presents the case for attending to clinical utility in the development and use of diagnostic criteria.

Discusses the importance of evaluating the value of assessment data in terms of their impact on the outcomes of psychological services.

Presents the case for why evidence-based assessment is needed in clinical psychology and some of the training-related challenges in ensuring the provision of evidence-based assessments.

- Barkham M, Margison F, Leach C, Lucock M, Mellor-Clark J, et al. 2001. Service profiling and outcomes benchmarking using the CORE-OM: toward practice-based evidence in the psychological therapies. *J. Consult. Clin. Psychol.* 69:184–96
- Bickman L, Rosof-Williams J, Salzer MS, Summerfelt WT, Noser K, et al. 2000. What information do clinicians value for monitoring adolescent client progress and outcomes? *Prof. Psychol. Res. Pract.* 31:70–74
- Blanton H, Jaccard J. 2006. Arbitrary metrics in psychology. *Am. Psychol.* 61:27–41
- Camara WJ, Nathan JS, Puente AE. 2000. Psychological test usage: implications in professional psychology. *Prof. Psychol. Res. Pract.* 31:141–54
- Cashel ML. 2002. Child and adolescent psychological assessment: current clinical practices and the impact of managed care. *Prof. Psychol. Res. Pract.* 33:446–53
- Charter RA. 2003. A breakdown of reliability coefficients by test type and reliability method, and the clinical implications of low reliability. *J. Gen. Psychol.* 130:290–304
- Crick NR, Nelson DA. 2002. Relational and physical victimization within friendships: Nobody told me there'd be friends like these. *J. Abnorm. Child Psychol.* 30:599–607
- De Los Reyes A, Kazdin AE. 2005. Informant discrepancies in the assessment of childhood psychopathology: a critical review, theoretical framework, and recommendations for further study. *Psychol. Bull.* 131:483–509**
- First MB, Pincus HA, Levine JB, Williams JBW, Ustun B, Peele R. 2004. Clinical utility as a criterion for revising psychiatric diagnoses. *Am. J. Psychiatry* 161:946–54**
- Flanagan DP, Kaufman AS. 2004. *Essentials of WISC-IV Assessment*. New York: Wiley
- Fletcher JM, Francis DJ, Morris RD, Lyon GR. 2005. Evidence-based assessment of learning disabilities in children and adolescents. *J. Clin. Child Adolesc. Psychol.* 34:506–22
- Garb HN. 2003. Incremental validity and the assessment of psychopathology in adults. *Psychol. Assess.* 15:508–20
- Garland AF, Kruse M, Aarons GA. 2003. Clinicians and outcome measurement: What's the use? *J. Behav. Health Serv. Res.* 30:393–405
- Geisinger KF. 1994. Cross-cultural normative assessment: translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychol. Assess.* 6:304–12
- Groth-Marnat G. 2003. *Handbook of Psychological Assessment*. Hoboken, NJ: Wiley. 4th ed.
- Hatfield DR, Ogles BM. 2004. The use of outcome measures by psychologists in clinical practice. *Prof. Psychol. Res. Pract.* 35:485–91
- Hayes SC, Nelson RO, Jarrett RB. 1987. The treatment utility of assessment: a functional approach to evaluating treatment quality. *Am. Psychol.* 42:963–74**
- Haynes SN, Lench HC. 2003. Incremental validity of new clinical assessment measures. *Psychol. Assess.* 15:456–66
- Heinemann AW. 2005. Putting outcome measurement in context: a rehabilitation psychology perspective. *Rehab. Psychol.* 50:6–14
- Horvath LS, Logan TK, Walker R. 2002. Child custody cases: a content analysis of evaluations in practice. *Prof. Psychol. Res. Pract.* 33:557–63
- Hunsley J. 2002. Psychological testing and psychological assessment: a closer examination. *Am. Psychol.* 57:139–40
- Hunsley J, Bailey JM. 1999. The clinical utility of the Rorschach: unfulfilled promises and an uncertain future. *Psychol. Assess.* 11:266–77
- Hunsley J, Bailey JM. 2001. Whither the Rorschach? An analysis of the evidence. *Psychol. Assess.* 13:472–85
- Hunsley J, Crabb R, Mash EJ. 2004. Evidence-based clinical assessment. *Clin. Psychol.* 57(3):25–32**

- Hunsley J, Lee CM, Wood J. 2003. Controversial and questionable assessment techniques. In *Science and Pseudoscience in Clinical Psychology*, ed. SO Lilienfeld, SJ Lynn, J Lohr, pp. 39–76. New York: Guilford
- Hunsley J, Mash EJ. 2005. Introduction to the special section on developing guidelines for the evidence-based assessment (EBA) of adult disorders. *Psychol. Assess.* 17:251–55
- Hunsley J, Mash EJ, eds. 2006. *A Guide to Assessments That Work*. New York: Oxford Univ. Press. In press
- Hunsley J, Meyer GJ. 2003. The incremental validity of psychological testing and assessment: conceptual, methodological, and statistical issues. *Psychol. Assess.* 15:446–55**
- Johnston C, Murray C. 2003. Incremental validity in the psychological assessment of children and adolescents. *Psychol. Assess.* 15:496–507
- Joiner TE, Walker RL, Pettit JW, Perez M, Cukrowicz KC. 2005. Evidence-based assessment of depression in adults. *Psychol. Assess.* 17:267–77
- Joneis T, Turkheimer E, Oltmanns TF. 2000. Psychometric analysis of racial differences on the Maudsley Obsessional Compulsive Inventory. *Assessment* 7:247–58
- Kaufman AS, Lichtenberger EO. 1999. *Essentials of WAIS-III Assessment*. New York: Wiley
- Kazdin AE. 2005. Evidence-based assessment for children and adolescents: issues in measurement development and clinical applications. *J. Clin. Child Adolesc. Psychol.* 34:548–58
- Kazdin AE. 2006. Arbitrary metrics: implications for identifying evidence-based treatments. *Am. Psychol.* 61:42–49
- Kendell R, Jablensky A. 2003. Distinguishing between the validity and utility of psychiatric diagnoses. *Am. J. Psychiatry* 160:4–12
- Kessler RC, Chiu WT, Demler O, Walters EE. 2005. Prevalence, severity, and Comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Arch. Gen. Psychiatry* 62:617–27
- Klein DN, Dougherty LR, Olino TM. 2005. Toward guidelines for evidence-based assessment of depression in children and adolescents. *J. Clin. Child Adolesc. Psychol.* 34:412–32
- Kraemer HC, Wilson KA, Hayward C. 2006. Lifetime prevalence and pseudocomorbidity in psychiatric research. *Arch. Gen. Psychiatry* 63:604–8
- Krueger RF, Chentsova-Dutton YE, Markon KE, Goldberg D, Ormel J. 2003. A cross-cultural study of the structure of comorbidity among common psychopathological syndromes in the general health care setting. *J. Abnorm. Psychol.* 112:437–47
- Krueger RF, Markon KE. 2006. Reinterpreting comorbidity: a model-based approach to understanding and classifying psychopathology. *Annu. Rev. Clin. Psychol.* 2:111–33
- Lally SJ. 2001. Should human figure drawings be admitted into the court? *J. Personal. Assess.* 76:135–49
- Lambert MJ, Whipple JL, Hawkins EJ, Vermeersch D, Nielsen SL, Smart DW. 2003. Is it time to track patient outcome on a routine basis? A meta-analysis. *Clin. Psychol. Sci. Pract.* 10:288–301**
- Lima EN, Stanley S, Kaboski B, Reitzel LR, Richey JA, et al. 2005. The incremental validity of the MMPI-2: When does therapist access not enhance treatment outcome? *Psychol. Assess.* 17:462–68
- Mash EJ, Barkley RA, eds. 2006. *Assessment of Childhood Disorders*. New York: Guilford. 4th ed. In press
- Mash EJ, Hunsley J. 2005. Evidence-based assessment of child and adolescent disorders: issues and challenges. *J. Clin. Child Adolesc. Psychol.* 34:362–79**
- MATRICES. 2006. Results of the MATRICES RAND panel meeting: average medians for the categories of each candidate test. <http://www.matrices.ucla.edu/matrices-psychometrics-frame.htm>

Provides an overview of numerous considerations in testing for, and using information about, incremental validity.

Presents meta-analytic data illustrating the value of treatment monitoring data in improving psychotherapy services.

Discusses key considerations in the development of guidelines for evidence-based assessment.

- McFall RM. 2005. Theory and utility—key themes in evidence-based assessment: comment on the special section. *Psychol. Assess.* 17:312–23
- McGrath RE. 2001. Toward more clinically relevant assessment research. *J. Personal. Assess.* 77:307–32
- McMahon RJ, Frick PJ. 2005. Evidence-based assessment of conduct problems in children and adolescents. *J. Clin. Child Adolesc. Psychol.* 34:477–505
- Meyer GJ, Archer RP. 2001. The hard science of Rorschach research: What do we know and where do we go? *Psychol. Assess.* 13:486–502
- Meyer GJ, Finn SE, Eyde L, Kay GG, Moreland KL, et al. 2001. Psychological testing and psychological assessment: a review of evidence and issues. *Am. Psychol.* 56:128–65
- Murray HA. 1943. *Thematic Apperception Test Manual*. Cambridge, MA: Harvard Univ. Press
- Nelson-Gray RO. 2003. Treatment utility of psychological assessment. *Psychol. Assess.* 15:521–31
- Newman DL, Moffitt TE, Caspi A, Silva PA. 1998. Comorbid mental disorders: implications for treatment and sample selection. *J. Abnorm. Psychol.* 107:305–11
- Nezu AM, McClure KS, Ronan GR, Meadows EA. 2000. *Practitioner's Guide to Empirically Based Measures of Depression*. Hingham, MA: Kluwer Plenum
- Nunnally JC, Bernstein IH. 1994. *Psychometric Theory*. New York: McGraw-Hill. 752 pp. 3rd ed.
- Ozonoff S, Goodlin-Jones BL, Solomon M. 2005. Evidence-based assessment of autism spectrum disorders in children and adolescents. *J. Clin. Child Adolesc. Psychol.* 34:523–40
- Pelham WE, Fabiano GA, Massetti GM. 2005. Evidence-based assessment of attention deficit hyperactivity disorder in children and adolescents. *J. Clin. Child Adolesc. Psychol.* 34:449–76
- Piotrowski C. 1999. Assessment practices in the era of managed care: current status and future directions. *J. Clin. Psychol.* 55:787–96
- Power TJ, Andrews TJ, Eiraldi RB, Doherty BJ, Ikeda MJ, et al. 1998. Evaluating attention deficit hyperactivity disorder using multiple informants: the incremental utility of combining teacher with parent reports. *Psychol. Assess.* 10:250–60
- Robinson JP, Shaver PR, Wrightsman LS. 1991. Criteria for scale selection and evaluation. In *Measures of Personality and Social Psychological Attitudes*, ed. JP Robinson, PR Shaver, LS Wrightsman, pp. 1–16. New York: Academic
- Rossini ED, Moretti RJ. 1997. Thematic Apperception Test (TAT) interpretation: practice recommendations from a survey of clinical psychology doctoral programs accredited by the American Psychological Association. *Prof. Psychol. Res. Pract.* 28:393–98
- Sackett DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS. 1996. Evidence based medicine: what it is and what it isn't. *Br. Med. J.* 312:71–72
- Schene AH, Koeter M, van Wijngaarden B, Knudsen HC, Leese M, et al. 2000. Methodology of a multi-site reliability study. *Br. J. Psychiatry* 177(Suppl. 39):15–20
- Schmidt FL, Le H, Ilies R. 2003. Beyond alpha: an empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychol. Methods* 8:206–24
- Sechrest L. 1963. Incremental validity: a recommendation. *Educ. Psychol. Meas.* 23:152–58
- Seto MC. 2005. Is more better? Combining actuarial risk scales to predict recidivism among adult sex offenders. *Psychol. Assess.* 17:156–67
- Silverman WK, Ollendick TH. 2005. Evidence-based assessment of anxiety and its disorders in children and adolescents. *J. Clin. Child Adolesc. Psychol.* 34:380–411
- Smith DA. 2002. Validity and values: monetary and otherwise. *Am. Psychol.* 57:136–37
- Snyder DK, Heyman RE, Haynes SN. 2005. Evidence-based approaches to assessing couple distress. *Psychol. Assess.* 17:288–307

- Spangler WD. 1992. Validity of questionnaire and TAT measures of need for achievement: two meta-analyses. *Psychol. Bull.* 112:140–54
- Streiner DL. 2003. Starting at the beginning: an introduction to coefficient alpha and internal consistency. *J. Personal. Assess.* 80:99–103
- Streiner DL, Norman GR. 2003. *Health Measurement Scales: A Practical Guide to Their Development and Use*. New York: Oxford Univ. Press. 283 pp. 3rd ed.
- Stricker G, Gold JR. 1999. The Rorschach: towards a nomothetically based, idiographically applicable configural model. *Psychol. Assess.* 11:240–50
- Vane JR. 1981. The Thematic Apperception Test: a review. *Clin. Psychol. Rev.* 1:319–36
- van Widenfelt BM, Treffers PDA, de Beurs E, Siebelink BM, Koudijs E. 2005. Translation and cross-cultural adaptation of assessment instruments used in psychological research with children and families. *Clin. Child Fam. Psychol. Rev.* 8:135–47
- Vermeersch DA, Lambert MJ, Burlingame GM. 2000. Outcome Questionnaire: item sensitivity to change. *J. Personal. Assess.* 74:242–61
- Vermeersch DA, Whipple JL, Lambert MJ, Hawkins EJ, Burchfield CM, Okiishi JC. 2004. Outcome Questionnaire: Is it sensitive to changes in counseling center clients? *J. Counsel. Psychol.* 51:38–49
- Watkins MW. 2003. IQ subtest analysis: clinical acumen or clinical illusion? *Sci. Rev. Mental Health Pract.* 2:118–41
- Weisz JR, Chu BC, Polo AJ. 2004. Treatment dissemination and evidence-based practice: strengthening intervention through clinician-researcher collaboration. *Clin. Psychol. Sci. Pract.* 11:300–7
- Widiger TA, Clark LA. 2000. Toward DSM-V and the classification of psychopathology. *Psychol. Bull.* 126:946–63
- Widiger TA, Samuel DB. 2005. Evidence-based assessment of personality disorders. *Psychol. Assess.* 17:278–87
- Wolraich ML, Lambert W, Doffing MA, Bickman L, Simmons T, Worley K. 2003. Psychometric properties of the Vanderbilt ADHD diagnostic parent rating scale in a referred population. *J. Pediatr. Psychol.* 28:559–68
- Wood JM, Garb HN, Lilienfeld SO, Nezworski MT. 2002. Clinical assessment. *Annu. Rev. Psychol.* 53:519–43
- Wood JM, Nezworski MT, Lilienfeld SO, Garb HN. 2003. *What's Wrong with the Rorschach?* San Francisco: Jossey-Bass
- Youngstrom EA, Findling RL, Calabrese JR, Gracious BL, Demeter C, et al. 2004. Comparing the diagnostic accuracy of six potential screening instruments for bipolar disorder in youths aged 5 to 17 years. *J. Am. Acad. Child Adolesc. Psychiatry* 43:847–58
- Youngstrom EA, Findling RL, Youngstrom JK, Calabrese JR. 2005. Toward an evidence-based assessment of pediatric bipolar disorder. *J. Clin. Child Adolesc. Psychol.* 34:433–48